Journal of Computer and System Sciences 76 (2010) 759–777

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Incentive compatible regression learning ☆

Ofer Dekel [a,1], Felix Fischer [b,*,2], Ariel D. Procaccia [b,3]

[a] Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
[b] Harvard SEAS, 33 Oxford Street, Cambridge, MA 02138, USA

ARTICLE INFO

Article history:
Received 14 February 2008
Received in revised form 1 March 2010
Available online 6 March 2010

Keywords:
Algorithmic mechanism design
Computational learning theory
Regression analysis

ABSTRACT

We initiate the study of incentives in a general machine learning framework. We focus on a game-theoretic regression learning setting where private information is elicited from multiple agents with different, possibly conflicting, views on how to label the points of an input space. This conflict potentially gives rise to untruthfulness on the part of the agents. In the restricted but important case when every agent cares about a single point, and under mild assumptions, we show that agents are motivated to tell the truth. In a more general setting, we study the power and limitations of mechanisms without payments. We finally establish that, in the general setting, the VCG mechanism goes a long way in guaranteeing truthfulness and economic efficiency.

1. Introduction

Machine learning is the area of computer science concerned with the design and analysis of algorithms that can learn from experience. A *supervised* learning algorithm observes a training set of labeled examples, and attempts to learn a rule that accurately predicts the labels of new examples. Following the rise of the Internet as a computational platform, machine learning problems have become increasingly dispersed, in the sense that different parts of the training set may be controlled by different computational or economic entities.

1.1. Motivation

Consider an Internet search company trying to improve the performance of their search engine by learning a ranking function from examples. The ranking function is the heart of a modern search engine, and can be thought of as a mapping that assigns a real-valued score to every pair of a query and a URL. Some of the large Internet search companies currently hire Internet users, whom we hereinafter refer to as "experts", to manually rank such pairs. These rankings are then pooled and used to train a ranking function. Moreover, the experts are chosen in a way such that averaging over the experts' opinions and interests presumably pleases the average Internet user.

However, different experts may have different interests and a different idea of the results a good search engine should return. For instance, take the ambiguous query "Jaguar", which has become folklore in search engine designer circles. The top

☆ An extended abstract of this paper appeared in the proceedings of the 19th Annual ACM–SIAM Symposium on Discrete Algorithms (SODA).

* Corresponding author. Fax: +1 617 495 8612.
E-mail addresses: oferd@microsoft.com (O. Dekel), fischerf@seas.harvard.edu (F. Fischer), arielpro@seas.harvard.edu (A.D. Procaccia).

[1] This work was done while the author was at The Hebrew University of Jerusalem.

[2] This work was done while the author was visiting The Hebrew University of Jerusalem and was supported by the School of Computer Science and Engineering, the Leibniz Center for Research in Computer Science, and by the Deutsche Forschungsgemeinschaft under grant BR 2312/3-1.

[3] This work was done while the author was at The Hebrew University of Jerusalem and was supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities.

0022-0000/$ – see front matter
doi:10.1016/j.jcss.2010.03.003

answer given by most search engines for this query is the website of the luxury car manufacturer. Knowing this, an animal-loving expert may decide to give this pair a disproportionately low score, hoping to improve the relative rank of websites dedicated to the Panthera Onca. An expert who is an automobile enthusiast may counter this measure by giving automotive websites a much higher score than is appropriate. From the search company's perspective, this type of strategic manipulation introduces an undesired bias in the training set.

As a second motivating example, consider the distribution process of a large retail chain, like Spanish fashion company Zara. Store managers typically report their predicted demand to the central warehouses, where global shipments of inventory are optimized. In recent years Zara has reengineered its distribution process using models from operations research [1,2]. In particular, regression learning is now employed to predict the upcoming weekly demand. The prediction is based on past sales data, but also on requests by store managers. This introduces incentives for the store managers, whose salaries depend largely on the sales in their own stores. Caro et al. [2] believe that "this caused store managers to frequently request quantities exceeding their true needs, particularly when they suspected that the warehouse might not hold enough inventory of a top-selling article to satisfy all stores. [...] Zara might in time consider introducing formal incentives for store managers to provide accurate forecasts, adding to its more traditional sales-related incentives." [2, p. 74].

1.2. Setting and goals

Our problem setting falls within the general boundaries of *statistical regression learning*. Regression learning is the task of constructing a real-valued function f based on a training set of examples, where each example consists of an input to the function and its corresponding output. In particular, the example $(\mathbf{x}, y)$ suggests that $f(\mathbf{x})$ should be equal to y. The accuracy of a function f on a given input–output pair $(\mathbf{x}, y)$ is defined using a loss function ℓ. Popular choices of the loss function are the squared loss, $\ell(f(\mathbf{x}), y) = (f(\mathbf{x}) - y)^2$, or the absolute loss, $\ell(f(\mathbf{x}), y) = |f(\mathbf{x}) - y|$. We typically assume that the training set is obtained by sampling i.i.d. from an underlying distribution over the product space of inputs and outputs. The overall quality of the function constructed by the learning algorithm is defined to be its expected loss, with respect to the same distribution.

We augment this well-studied setting by introducing a set of *strategic agents*. Each agent holds as private information an individual distribution over the input space and values for the points in the support of this distribution, and measures the quality of a regression function with respect to this data. The global goal, on the other hand, is to do well with respect to the average of the individual points of view. A training set is obtained by eliciting private information from the agents, who may reveal this information untruthfully in order to favorably influence the result of the learning process.

Mechanism design is a subfield of economics that is concerned with the question of how to incentivize agents to truthfully report their private information, also known as their type. Given potentially non-truthful reports from the agents, a mechanism determines a global solution, and possibly additional monetary transfers to and from the agents. A mechanism is said to be *incentive compatible* if it is always in the agents' best interest to report their true types, and *efficient* if the solution maximizes social welfare (i.e., minimizes the overall loss). Our goal in this paper will be to design and analyze incentive compatible and efficient mechanisms for the regression learning setting. It should be noted that incentive compatibility is essential for obtaining *any* learning theoretic bounds. Otherwise, all agents might reveal untruthful information at the same time, in a coordinated or uncoordinated way, causing the learning problem itself to be ill-defined.

1.3. Results

We begin our investigation by considering a restricted setting where each agent is only interested in a single point of the input space. Quite surprisingly, it turns out that a specific choice of ℓ, namely the absolute loss function, leads to excellent game-theoretic properties: an algorithm which simply finds an empirical risk minimizer on the training set is group strategyproof, meaning that no coalition of agents is motivated to lie. Like in all of our incentive compatibility results, truthfulness holds with respect to dominant strategies, i.e., regardless of the other agents' actions. In a sense, this is the strongest incentive property that could possibly be obtained. We also show that even much weaker truthfulness results cannot be obtained for a wide range of other loss functions, including the popular squared loss.

In the more general case where agents are interested in non-degenerate distributions, achieving incentive compatibility requires more sophisticated mechanisms. We show that the well-known VCG mechanism does very well: with probability $1 - \delta$, no agent can gain more than ϵ by lying, where both ϵ and δ can be made arbitrarily small by increasing the size of the training set. This result holds for any choice of loss function ℓ.

We also study what happens when payments are disallowed. In this setting, we obtain limited positive results for the absolute loss function and for restricted yet interesting function classes. In particular, we present a mechanism which is approximately group strategyproof as above and 3-efficient in the sense that the solution provides a 3-approximation to optimal social welfare. We complement these results with a matching lower bound and provide strong evidence that no approximately incentive compatible and approximately efficient mechanism exists for more expressive functions classes.

1.4. Related work

To the best of our knowledge, this paper is the first to study incentives in a general machine learning framework. Previous work in machine learning has investigated the related problem of learning in the presence of inconsistent and noisy training

data, where the noise can be either random [3,4] or adversarial [5,6]. Barreno et al. [7] consider a specific situation where machine learning is used as a component of a computer security system, and account for the possibility that the training data is subject to a strategic attack intended to infiltrate the secured system. In contrast to these approaches, we do not attempt to design algorithms that can tolerate noise, but instead focus on designing algorithms that discourage the strategic addition of noise.

Most closely related to our work is that of Perote and Perote-Peña [8]. The authors essentially study the setting where each agent controls one point of the input space, in a framework that is not learning-theoretic. In addition, they only consider linear regression, and the input space is restricted to be the real line. For that setting, the authors put forward a class of truthful estimators. Rather than looking at the approximation properties of said estimators, they are instead shown to be Pareto-optimal, i.e., there exist no regression lines that are weakly better for all agents, and strictly better for at least one agent.

Our work is also related to the area of *algorithmic mechanism design*, introduced in the seminal work of Nisan and Ronen [9]. Algorithmic mechanism design studies algorithmic problems in a game-theoretic setting where the different participants cannot be assumed to follow the algorithm but rather act in a selfish way. It has turned out that the main challenge of algorithmic mechanism design is the inherent incompatibility of generic truthful mechanisms with approximation schemes for hard algorithmic problems. As a consequence, most of the current work in algorithmic mechanism design focuses on dedicated mechanisms for hard problems (see, e.g., [10,11]). What distinguishes our setting from that of algorithmic mechanism design is the need for *generalization* to achieve globally satisfactory results on the basis of a small number of samples. Due to the dynamic and uncertain nature of the domain, inputs are usually assumed to be drawn from some underlying fixed distribution. The goal then is to design algorithms that, with high probability, perform well on samples drawn from the same distribution.

More distantly related to our work is research which applies machine learning techniques in game theory and mechanism design. Balcan et al. [12], for instance, use techniques from sample complexity to reduce mechanism design problems to standard algorithmic problems. Another line of research puts forward that machine learning can be used to predict consumer behavior, or find a concise description for collective decision making. Work along this line includes the learnability of choice functions and choice correspondences [13,14].

1.5. Structure of the paper

In the following section, we introduce the necessary concepts from mechanism design. In Section 3, we give a general exposition of regression learning and introduce our model of regression learning with multiple agents. We then examine three settings of increasing generality: in Section 4, we consider the case where the distribution of each agent puts all of the weight on a single point of the input space; in Section 5, we then move to the more general setting where the distribution of each agent is a discrete distribution supported on a finite set of points; we finally investigate arbitrary distributions in Section 6, leveraging the results of the previous sections. In Section 7, we discuss our results and give some directions for future research.

2. Preliminaries

A mechanism design problem (see, e.g., [15]) is given by a set $N = \{1, 2, \ldots, n\}$ of *agents* that interact to select one element from a set A of alternatives. Agent $i \in N$ is associated with a *type* θ_i from a set Θ_i of possible types, corresponding to the private information held by this agent. We write $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ for a profile of types for the different agents and $\Theta = \prod_{i \in N} \Theta_i$ for the set of possible type profiles. $\theta_{-i} \in \Theta_{-i}$ is used to denote a profile of types for all agents but i. Furthermore, agent $i \in N$ employs *preferences* over A, represented by a real-valued *valuation function* $v_i : A \times \Theta_i \to \mathbb{R}$. In this paper, we only consider settings of *private values* where an agent's preferences depend exclusively on his type.

A *social choice function* is a function $f : \Theta \to A$. One desirable property of social choice functions is *efficiency*. A social choice function f is called α-efficient if for all $\theta \in \Theta$,

$$\alpha \cdot \sum_{i \in N} v_i\big(f(\theta), \theta_i\big) \geqslant \max_{a \in A} \sum_{i \in N} v_i(a, \theta_i).$$

We say that a social choice function is efficient if it is 1-efficient and approximately efficient if it is α-efficient for some α.

Agents' types, and thus the input to f, are private, and agents may strategically report information that does not agree with their true type in order to increase their payoff at the expense of social welfare. The goal of mechanism design is to provide incentives to the agents to report their true types and enable the computation of a socially optimal solution. In order to achieve this, it may sometimes be necessary to tax or subsidize the different agents based on their revealed type. This is done by means of a *payment function* $p : \Theta \to \mathbb{R}^n$. Intuitively, $p_i(\theta)$ represents a payment from agent i to the mechanism if the revealed types are θ.

As this definition indicates, we will restrict our attention to the class of *direct revelation mechanisms*, where all agents simultaneously announce their types within a single round. We will see momentarily that this does not imply a restriction in expressiveness with respect to the problems studied in this paper. Formally, a (direct revelation) mechanism is a pair (f, p) of a social choice function f and a payment function p. A mechanism (f, p) will be called α-efficient if f is α-efficient.

Game-theoretic reasoning, more specifically a model of strict incomplete information games, is used to analyze how agents interact with a mechanism, a desirable criterion being stability according to some game-theoretic solution concept. Each agent $i \in N$ has a true type $\theta_i \in \Theta_i$, and reveals some type $\hat{\theta}_i \in \Theta_i$. Agents have no information, not even distributional, about the types of the other agents. A *strategy* is a function mapping true types to revealed types, and an outcome is chosen according to the profile of revealed types. The *payoff* of an agent thus depends on his true type and the revealed types of all the agents.

Let $\epsilon \geqslant 0$. A mechanism f is said to be *ϵ-group strategyproof* (in dominant strategy equilibrium) if for any coalition $C \subseteq N$ of the agents, the only way that all members of C can gain at least ϵ by jointly deviating from the profile θ of true types is for all of them to gain exactly ϵ. More formally, consider $\hat{\theta} \in \Theta$ such that $\hat{\theta}_j = \theta_j$ whenever $j \notin C$. Then, ϵ-group strategyproofness requires that if for all $i \in C$,

$$v_i\big(f(\hat{\theta}), \theta_i\big) - p_i(\hat{\theta}) \geqslant v_i\big(f(\theta), \theta_i\big) - p_i(\theta) + \epsilon,$$

then for all $i \in C$,

$$v_i\big(f(\hat{\theta}), \theta_i\big) - p_i(\hat{\theta}) = v_i\big(f(\theta), \theta_i\big) - p_i(\theta) + \epsilon.$$

A mechanism is called *ϵ-strategyproof* if the above is satisfied for any $C \subseteq N$ such that $|C| = 1$. We then say that a mechanism is (group) strategyproof if it is 0-(group) strategyproof. In other words, group strategyproofness requires that if some member of an arbitrary coalition of agents strictly gains from a joint deviation by the coalition, then some other member must strictly lose. A social choice function will sometimes be referred to as a mechanism (without payments) if the distinction is obvious from the context. A social choice function f is then called (group) strategyproof if the mechanism (f, p^0) is (group) strategyproof, where p^0 is the constant zero function.

Strategyproofness is sometimes defined in a way that includes individual rationality, and the term incentive compatibility is then reserved for the above property that agents cannot gain by revealing their types untruthfully. We do not make such a distinction in this paper but rather use the terms incentive compatibility, truthfulness, and strategyproofness interchangeably. We note two things, however. First, individual rationality is trivially satisfied in our case by any mechanism without payments, as will become apparent later. Secondly, it is not immediately clear how to achieve individual rationality for mechanisms with payments.

If we say that a mechanism is *not* strategyproof, we mean it is not strategyproof in the weaker solution concept of (ex-post) Nash equilibrium, i.e., there exists a strategy profile under which some agent can gain from untruthful revelation, even if all other agents are assumed to reveal their types truthfully. Due to the well-known *revelation principle*, only direct mechanisms need to be considered in order to answer the question of whether there exists a mechanism that is incentive compatible in dominant strategy or Nash equilibrium.

We conclude this section with a general mechanism due to Vickrey [16], Clarke [17], and Groves [18]. This mechanism starts from an efficient social choice function f and computes each agent's payment according to the social welfare of the other agents, thus aligning his interests with that of society. Formally, a mechanism (f, p) is called Vickrey–Clarke–Groves (VCG) mechanism if f is efficient and there exist functions $h_i : \Theta_{-i} \to \mathbb{R}$ such that

$$p_i(\theta) = h_i(\theta_{-i}) - \sum_{j \neq i} v_j\big(f(\theta), \theta_j\big).$$

VCG mechanisms are strategyproof [18] but in general not group strategyproof. The latter is due to the fact that in some cases the members of a coalition can influence each others' payments such that all of them gain. Interestingly, all mechanisms with unrestricted type spaces that are efficient and strategyproof (in dominant strategy equilibrium) are VCG mechanisms.

3. The model

In this section we formalize the regression learning problem described in the introduction and cast it in the framework of game theory. Some of the definitions are illustrated by relating them to the Internet search example presented in Section 1.

We focus on the task of learning a real-valued function over an *input space* $\mathcal{X}$. In the Internet search example, $\mathcal{X}$ would be the set of all query-URL pairs, and our task would be to learn the ranking function of a search engine. Let $N = \{1, \ldots, n\}$ be a set of agents, which in our running example would be the set of all experts. For each agent $i \in N$, let o_i be a function from $\mathcal{X}$ to $\mathbb{R}$ and let ρ_i be a probability distribution over $\mathcal{X}$. Intuitively, o_i is what agent i thinks to be the correct real-valued function, while ρ_i captures the relative importance that agent i assigns to different parts of $\mathcal{X}$. In the Internet search example, o_i would be the optimal ranking function according to agent i, and ρ_i would be a distribution over query-URL pairs that assigns higher weight to queries from that agent's areas of interest.

Let $\mathcal{F}$ be a class of functions, where every $f \in \mathcal{F}$ is a function from $\mathcal{X}$ to the real line. We call $\mathcal{F}$ the *hypothesis space* of our problem, and restrict the output of the learning algorithm to functions in $\mathcal{F}$. We evaluate the accuracy of each $f \in \mathcal{F}$ using a *loss function* $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$. For a particular input–output pair $(\mathbf{x}, y)$, we interpret $\ell(f(\mathbf{x}), y)$ as the penalty associated with predicting the output value $f(\mathbf{x})$ when the true output is known to be y. As mentioned in the introduction,

common choices of ℓ are the squared loss, $\ell(\alpha,\beta)=(\alpha-\beta)^2$, and the absolute loss, $\ell(\alpha,\beta)=|\alpha-\beta|$. The accuracy of a hypothesis $f\in\mathcal{F}$ is defined to be the average loss of f over the entire input space. Formally, define the *risk* associated by agent i with the function f as

$$R_i(f)=\mathbb{E}_{\mathbf{x}\sim\rho_i}\big[\ell\big(f(\mathbf{x}),o_i(x)\big)\big].$$

Clearly, this subjective definition of hypothesis accuracy allows for different agents to have significantly different valuations of different functions in $\mathcal{F}$, and it is quite possible that we will not be able to please all of the agents simultaneously. Instead, our goal is to satisfy the agents in N on average. Define J to be a random variable distributed uniformly over the elements of N. Now define the *global risk* of a function f to be the average risk with respect to all of the agents, namely

$$R_N(f)=\mathbb{E}\big[R_J(f)\big].$$

We are now ready to state our learning-theoretic goal formally: we would like to find a hypothesis in $\mathcal{F}$ that attains a global risk as close as possible to $\inf_{f\in\mathcal{F}}R_N(F)$.

Even if N is small, we still have no explicit way of calculating $R_N(f)$. Instead, we use an empirical estimate of the risk as a proxy to the risk itself. For each $i\in N$, we randomly sample m points independently from the distribution ρ_i and request their respective labels from agent i. In this way, we obtain the labeled training set $\tilde{S}_i=\{(\mathbf{x}_{i,j},\tilde{y}_{i,j})\}_{j=1}^m$. Agent i may label the points in $\tilde{S}_i$ however he sees fit, and we therefore say that agent i *controls* (the labels of) these points. We usually denote agent i's "true" training set by $S_i=\{(\mathbf{x}_{ij},y_{ij})\}_{j=1}^m$, where $y_{ij}=o_i(x_{ij})$. After receiving labels from all agents in N, we define the *global training set* to be the multiset $\tilde{S}=\biguplus_{i\in N}\tilde{S}_i$.

The elicited training set $\tilde{S}$ is presented to a regression learning algorithm, which in return constructs a *hypothesis* $\tilde{f}\in\mathcal{F}$. Each agent can influence $\tilde{f}$ by modifying the labels he controls. This observation brings us to the game-theoretic aspect of our setting. For all $i\in N$, agent i's private information, or type, is a vector of true labels $y_{ij}=o_i(x_{ij})$, $j=1,\ldots,m$. The sampled points $\mathbf{x}_{ij}$, $j=1,\ldots,m$, are exogenously given and assumed to be common knowledge. The *strategy space* of each agent then consists of all possible *values* for the labels he controls. In other words, agent i reports a labeled training set $\tilde{S}_i$. We sometimes use $\tilde{S}_{-i}$ as a shorthand for $\tilde{S}\setminus\tilde{S}_i$, the strategy profile of all agents except agent i. The space of possible outcomes is the hypothesis space $\mathcal{F}$, and the utility of agent i for an outcome $\tilde{f}$ is determined by his risk $R_i(\tilde{f})$. More precisely, agent i chooses $\tilde{y}_{i1},\ldots,\tilde{y}_{im}$ so as to minimize $R_i(f)$. We follow the usual game-theoretic assumption that he does this with full knowledge of the inner workings of our regression learning algorithm, and name the resulting game the *learning game*.

Notice that under the above formalism, a regression learning algorithm is in fact a social choice function, which maps the types of the agents to a hypothesis. One of the simplest and most popular regression learning techniques is *empirical risk minimization* (ERM). The *empirical risk* associated with a hypothesis f, with respect to a sample S, is denoted by $\hat{R}(f,S)$ and defined to be the average loss attained by f on the examples in S, i.e.,

$$\hat{R}(f,S)=\frac{1}{|S|}\sum_{(\mathbf{x},y)\in S}\ell\big(f(\mathbf{x}),y\big).$$

An ERM algorithm finds the empirical risk minimizer $\hat{f}$ within $\mathcal{F}$. More formally,

$$\hat{f}=\operatorname*{arg\,min}_{f\in\mathcal{F}}\hat{R}(f,S).$$

A large part of this paper will be dedicated to ERM algorithms. For some choices of loss function and hypothesis class, it may occur that the global minimizer of the empirical risk is not unique, and we must define an appropriate tie-breaking mechanism.

Since our strategy is to use $\hat{R}(f,\tilde{S})$ as a surrogate for $R_N(f)$, we need $\hat{R}(f,\tilde{S})$ to be an unbiased estimator of $R_N(f)$. A particular situation in which this can be achieved is when all agents $i\in N$ truthfully report $\tilde{y}_{ij}=o_i(\mathbf{x}_{ij})$ for all j. It is important to note that truthfulness need not come at the expense of the overall solution quality. This can be seen by a variation of the well-known revelation principle already mentioned in Section 2. Assume that for a given mechanism and given true inputs there is an equilibrium in which some agents report their inputs untruthfully, and which leads to an outcome that is strictly better than any outcome achievable by an incentive compatible mechanism. Then we can design a new mechanism that, given the true inputs, simulates the agents' lies and yields the exact same output in equilibrium.

4. Degenerate distributions

We begin our study by focusing on a special case, where each agent is only interested in a single point of the input space. Even this simple setting has interesting applications. Consider for example the problem of allocating tasks among service providers, e.g., messages to routers, jobs to remote processors, or reservations of bandwidth to Internet providers. Machine learning techniques are used to obtain a global picture of the capacities, which in turn are private information of the respective providers. Regression learning provides an appropriate model in this context, as each provider is interested

in an allocation that is as close as possible to its capacity: more tasks mean more revenue, but an overload is clearly undesirable.

A concrete economic motivation for this setting is given by Perote and Perote-Peña [8]. The authors consider a monopolist trade union in some sector that has to set a common hourly wage for its members. The union collects information about the hours of work in each firm versus the firm's expected profitability, and accordingly sets a single sectorial wage per hour. The hours of work are public information, but the expected profitability is private. Workers that are more profitable might have an incentive to exaggerate their profitability in order to increase the hourly common wage.

More formally, the distribution ρ_i of agent i is now assumed to be degenerate, and the sample S_i becomes a singleton. Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote the set of true input–output pairs, where now $y_i = o_i(\mathbf{x}_i)$, and $S_i = \{(\mathbf{x}_i, y_i)\}$ is the single example controlled by agent i. Each agent selects an output value $\tilde{y}_i$, and the reported (possibly untruthful) training set $\tilde{S} = \{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^n$ is presented to a regression learning algorithm. The algorithm constructs a hypothesis $\tilde{f}$ and agent i's cost is the loss

$$R_i(\tilde{f}) = \mathbb{E}_{\mathbf{x}\sim\rho_i}\big[\ell\big(\tilde{f}(\mathbf{x}), o_i(\mathbf{x})\big)\big] = \ell\big(\tilde{f}(\mathbf{x}_i), y_i\big)$$

on the point he controls, where ℓ is a predefined loss function. Within this setting, we examine the game-theoretic properties of ERM.

As noted above, an ERM algorithm takes as input a loss function ℓ and a training set S, and outputs the hypothesis that minimizes the empirical risk on S according to ℓ. Throughout this section, we write $\hat{f} = \mathrm{ERM}(\mathcal{F}, \ell, S)$ as shorthand for $\arg\min_{f\in\mathcal{F}} \hat{R}(f, \ell, S)$. We restrict our discussion to loss functions of the form $\ell(\alpha, \beta) = \mu(|\alpha - \beta|)$, where $\mu : \mathbb{R}_+ \to \mathbb{R}$ is a monotonically increasing convex function, and to the case where $\mathcal{F}$ is a convex set of functions. These assumptions enable us to cast ERM as a convex optimization problem, which are typically tractable. Most choices of ℓ and $\mathcal{F}$ that do not satisfy the above constraints may not allow for computationally efficient learning, and are therefore less interesting.

We prove two main theorems: if μ is a linear function, then ERM is group strategyproof; if on the other hand μ grows faster than any linear function, and given minimal conditions on $\mathcal{F}$, ERM is not strategyproof.

4.1. ERM with the absolute loss

In this section, we focus on the absolute loss function. Indeed, let ℓ denote the absolute loss, $\ell(a, b) = |a - b|$, and let $\mathcal{F}$ be a convex hypothesis class. Because ℓ is only weakly convex, there may be multiple hypotheses in $\mathcal{F}$ that globally minimize the empirical risk and we must add a tie-breaking step to our ERM algorithm. Concretely, consider the following two-step procedure:

1. Empirical risk minimization: calculate
$$r = \min_{f\in\mathcal{F}} \hat{R}(f, S).$$
2. Tie-breaking: return
$$\tilde{f} = \operatorname*{arg\,min}_{f\in\mathcal{F}:\ \hat{R}(f,S)=r} \|f\|,$$
where $\|f\|^2 = \int f^2(\mathbf{x})\, d\mathbf{x}$.

Our assumption that $\mathcal{F}$ is a convex set implies that the set of empirical risk minimizers $\{f \in \mathcal{F}:\ \hat{R}(f, S) = r\}$ is also convex. The function $\|f\|$ is a strictly convex function and therefore the output of the tie-breaking step is uniquely defined.

For example, imagine that $\mathcal{X}$ is the unit ball in $\mathbb{R}^n$ and that $\mathcal{F}$ is the set of homogeneous linear functions, of the form $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x}\rangle$, where $\mathbf{w} \in \mathbb{R}^n$. In this case, Step 1 above can be restated as the following linear program:

$$\min_{\boldsymbol{\xi}\in\mathbb{R}^m, \mathbf{w}\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^m \xi_i \quad \text{s.t.} \quad \forall i \quad \langle \mathbf{w}, \mathbf{x}_i\rangle - y_i \leqslant \xi_i \quad \text{and} \quad y_i - \langle \mathbf{w}, \mathbf{x}_i\rangle \leqslant \xi_i.$$

The tie-breaking step can then be written as the following quadratic program with linear constraints:

$$\operatorname*{arg\,min}_{\boldsymbol{\xi}\in\mathbb{R}^m, \mathbf{w}\in\mathbb{R}^n} \|\mathbf{w}\|^2 \quad \text{s.t.} \quad \sum_{i=1}^m \xi_i = r \quad \text{and} \quad \forall i \quad \langle \mathbf{w}, \mathbf{x}_i\rangle - y_i \leqslant \xi_i \quad \text{and} \quad y_i - \langle \mathbf{w}, \mathbf{x}_i\rangle \leqslant \xi_i.$$

In our analysis, we only use the fact that $\|f\|$ is a strictly convex function of f. Any other strictly convex function can be used in its place in the tie-breaking step.

The following theorem states that ERM using the absolute loss function has excellent game-theoretic properties. More precisely, it is group strategyproof: if a member of an arbitrary coalition of agents strictly gains from a joint deviation by the coalition, then some other member must strictly lose. It should also be noted that in our case any mechanism without

payments satisfies individual rationality: if some agent does not provide values for his part of the sample, then ERM will simply return the best fit for the points of the other agents, so no agent can gain by not taking part in the mechanism.

Theorem 4.1. *Let N be a set of agents, $S = \biguplus_{i\in N} S_i$ a training set such that $S_i = \{\mathbf{x}_i, y_i\}$ for all $i \in N$, and let ρ_i be degenerate at $\mathbf{x}_i$. Let ℓ denote the absolute loss, $\ell(a,b) = |a-b|$, and let $\mathcal{F}$ be a convex hypothesis class. Then, ERM minimizing ℓ over $\mathcal{F}$ with respect to S is group strategyproof.*

We prove this theorem below, as a corollary of the following more explicit result.

Proposition 4.2. *Let $\hat{S} = \{(\mathbf{x}_i, \hat{y}_i)\}_{i=1}^m$ and $\tilde{S} = \{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^m$ be two training sets on the same set of points, and let $\hat{f} = \mathtt{ERM}(\mathcal{F}, \ell, \hat{S})$ and $\tilde{f} = \mathtt{ERM}(\mathcal{F}, \ell, \tilde{S})$. If $\hat{f} \neq \tilde{f}$ then there exists $i \in N$ such that $\hat{y}_i \neq \tilde{y}_i$ and $\ell(\hat{f}(\mathbf{x}_i), \hat{y}_i) < \ell(\tilde{f}(\mathbf{x}_i), \hat{y}_i)$.*

Proof. Let U be the set of indices on which $\hat{S}$ and $\tilde{S}$ disagree, i.e., $U = \{i\colon\ \hat{y}_i \neq \tilde{y}_i\}$. We prove the claim by proving its counter-positive, i.e., we assume that $\ell(\tilde{f}(\mathbf{x}_i), \hat{y}_i) \leqslant \ell(\hat{f}(\mathbf{x}_i), \hat{y}_i)$ for all $i \in U$, and prove that $\hat{f} \equiv \tilde{f}$. We begin by considering functions of the form $f_\alpha(\mathbf{x}) = \alpha\tilde{f}(\mathbf{x}) + (1-\alpha)\hat{f}(\mathbf{x})$ and proving that there exists $\alpha \in (0,1]$ for which

$$\hat{R}(\hat{f}, \tilde{S}) - \hat{R}(\hat{f}, \hat{S}) = \hat{R}(f_\alpha, \tilde{S}) - \hat{R}(f_\alpha, \hat{S}). \tag{1}$$

For every $i \in U$, our assumption that $\ell(\tilde{f}(\mathbf{x}_i), \hat{y}_i) \leqslant \ell(\hat{f}(\mathbf{x}_i), \hat{y}_i)$ implies that one of the following four inequalities holds:

$$\tilde{f}(\mathbf{x}_i) \leqslant \hat{y}_i < \hat{f}(\mathbf{x}_i), \qquad \tilde{f}(\mathbf{x}_i) \geqslant \hat{y}_i > \hat{f}(\mathbf{x}_i), \tag{2}$$

$$\hat{y}_i \leqslant \tilde{f}(\mathbf{x}_i) \leqslant \hat{f}(\mathbf{x}_i), \qquad \hat{y}_i \geqslant \tilde{f}(\mathbf{x}_i) \geqslant \hat{f}(\mathbf{x}_i). \tag{3}$$

Furthermore, we assume without loss of generality that $\tilde{y}_i = \tilde{f}(\mathbf{x}_i)$ for all $i \in U$. Otherwise, we could simply change $\tilde{y}_i$ to equal $\tilde{f}(\mathbf{x}_i)$ for all $i \in U$ without changing the output of the learning algorithm. If one of the two inequalities in (2) holds, we set

$$\alpha_i = \frac{\hat{y}_i - \hat{f}(\mathbf{x}_i)}{\tilde{f}(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)},$$

and note that $\alpha_i \in (0,1]$ and $f_{\alpha_i}(\mathbf{x}_i) = \hat{y}_i$. Therefore, for every $\alpha \in (0, \alpha_i]$ it holds that either

$$\tilde{y}_i \leqslant \hat{y}_i \leqslant f_\alpha(\mathbf{x}_i) < \hat{f}(\mathbf{x}_i) \quad \text{or} \quad \tilde{y}_i \geqslant \hat{y}_i \geqslant f_\alpha(\mathbf{x}_i) > \hat{f}(\mathbf{x}_i).$$

Setting $c_i = |\hat{y}_i - \tilde{y}_i|$, we conclude that for all α in $(0, \alpha_i]$,

$$\ell\big(\hat{f}(\mathbf{x}_i), \tilde{y}_i\big) - \ell\big(\hat{f}(\mathbf{x}_i), \hat{y}_i\big) = c_i \quad \text{and} \quad \ell\big(f_\alpha(\mathbf{x}_i), \tilde{y}_i\big) - \ell\big(f_\alpha(\mathbf{x}_i), \hat{y}_i\big) = c_i. \tag{4}$$

Alternatively, if one of the inequalities in (3) holds, we have that either

$$\hat{y}_i \leqslant \tilde{y}_i \leqslant f_\alpha(\mathbf{x}_i) \leqslant \hat{f}(\mathbf{x}_i) \quad \text{or} \quad \hat{y}_i \geqslant \tilde{y}_i \geqslant f_\alpha(\mathbf{x}_i) \geqslant \hat{f}(\mathbf{x}_i).$$

Setting $\alpha_i = 1$ and $c_i = -|\tilde{y}_i - \hat{y}_i|$, we once again have that (4) holds for all α in $(0, \alpha_i]$. Moreover, if we choose $\alpha = \min_{i\in U} \alpha_i$, (4) holds simultaneously for all $i \in U$. (4) also holds trivially for all $i \notin U$ with $c_i = 0$. (1) can now be obtained by summing both of the equalities in (4) over all i.

Next, we recall that $\mathcal{F}$ is a convex set and therefore $f_\alpha \in \mathcal{F}$. Since $\hat{f}$ minimizes the empirical risk with respect to $\hat{S}$ over $\mathcal{F}$, we specifically have that

$$\hat{R}(\hat{f}, \hat{S}) \leqslant \hat{R}(f_\alpha, \hat{S}). \tag{5}$$

Combining this inequality with (1) results in

$$\hat{R}(\hat{f}, \tilde{S}) \leqslant \hat{R}(f_\alpha, \tilde{S}). \tag{6}$$

Since the empirical risk function is convex in its first argument, we have that

$$\hat{R}(f_\alpha, \tilde{S}) \leqslant \alpha\hat{R}(\tilde{f}, \tilde{S}) + (1-\alpha)\hat{R}(\hat{f}, \tilde{S}). \tag{7}$$

Replacing the left-hand side above with its lower bound in (6) yields $\hat{R}(\hat{f}, \tilde{S}) \leqslant \hat{R}(\tilde{f}, \tilde{S})$. On the other hand, we know that $\tilde{f}$ minimizes the empirical risk with respect to $\tilde{S}$, and specifically $\hat{R}(\tilde{f}, \tilde{S}) \leqslant \hat{R}(\hat{f}, \tilde{S})$. Overall, we have shown that

$$\hat{R}(\hat{f}, \tilde{S}) = \hat{R}(\tilde{f}, \tilde{S}) = \min_{f\in\mathcal{F}} \hat{R}(f, \tilde{S}). \tag{8}$$

Next, we turn our attention to $\|\hat{f}\|$ and $\|\tilde{f}\|$. We start by combining (8) with (7) to get $\hat{R}(f_\alpha, \tilde{S}) \leqslant \hat{R}(\hat{f}, \tilde{S})$. Recalling (1), we have that $\hat{R}(f_\alpha, \hat{S}) \leqslant \hat{R}(\hat{f}, \hat{S})$. Once again using (5), we conclude that $\hat{R}(f_\alpha, \hat{S}) = \hat{R}(\hat{f}, \hat{S})$. Although $\hat{f}$ and f_α both minimize the empirical risk with respect to $\hat{S}$, we know that $\hat{f}$ was chosen as the output of the algorithm, and therefore it must hold that

$$\|\hat{f}\| \leqslant \|f_\alpha\|. \tag{9}$$

Using convexity of the norm, we have $\|f_\alpha\| \leqslant \alpha\|\tilde{f}\| + (1-\alpha)\|\hat{f}\|$. Combining this inequality with (9), we get $\|\hat{f}\| \leqslant \|\tilde{f}\|$. On the other hand, (8) tells us that both $\hat{f}$ and $\tilde{f}$ minimize the empirical risk with respect to $\tilde{S}$, whereas $\tilde{f}$ is chosen as the algorithm output, so $\|\tilde{f}\| \leqslant \|\hat{f}\|$. Overall, we have shown that

$$\|\hat{f}\| = \|\tilde{f}\| = \min_{f \in \mathcal{F}:\ \hat{R}(f,\tilde{S}) = \hat{R}(\tilde{f},\tilde{S})} \|f\|. \tag{10}$$

In summary, in (8) we showed that both $\hat{f}$ and $\tilde{f}$ minimize the empirical risk with respect to $\tilde{S}$, and therefore both move on to the tie breaking step of the algorithm. Then, in (10) we showed that both functions attain the minimum norm over all empirical risk minimizers. Since the norm is strictly convex, its minimum is unique, and therefore $\hat{f} \equiv \tilde{f}$. □

To understand the intuition behind Proposition 4.2, as well as its relation to Theorem 4.1, assume that $\hat{S}$ represents the true preferences of the agents, and that $\tilde{S}$ represents the values revealed by the agents and used to train the regression function. Moreover, assume that $\hat{S} \neq \tilde{S}$. Proposition 4.2 states that one of two things can happen. Either $\hat{f} \equiv \tilde{f}$, i.e., revealing the values in $\tilde{S}$ instead of the true values in $\hat{S}$ does not affect the result of the learning process. In this case, the agents might as well have told the truth. Or, $\hat{f}$ and $\tilde{f}$ are different hypotheses, and Proposition 4.2 tells us that there must exist an agent i who lied about his true value and is strictly worse off due to his lie. Clearly, agent i has no incentive to actually participate in such a lie. This said, we can now proceed to prove the theorem.

Proof of Theorem 4.1. Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ be a training set that represents the true private information of a set N of agents and let $\tilde{S} = \{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^m$ be the information revealed by the agents and used to train the regression function. Let $C \subseteq N$ be an arbitrary coalition of agents that have conspired to decrease some of their respective losses by lying about their values. Now define the hybrid set of values where

$$\text{for all } i \in N, \quad \hat{y}_i = \begin{cases} y_i & \text{if } i \in C, \\ \tilde{y}_i & \text{otherwise,} \end{cases}$$

and let $\hat{S} = \{(\mathbf{x}_i, \hat{y}_i)\}_{i=1}^m$. Finally, let $\hat{f} = \texttt{ERM}(\mathcal{F}, \ell, \hat{S})$ and $\tilde{f} = \texttt{ERM}(\mathcal{F}, \ell, \tilde{S})$.

If $\hat{f} \equiv \tilde{f}$ then the members of C gain nothing from being untruthful. Otherwise, Proposition 4.2 states that there exists an agent $i \in N$ such that $\hat{y}_i \neq \tilde{y}_i$ and $\ell(\hat{f}(\mathbf{x}_i), \hat{y}_i) < \ell(\tilde{f}(\mathbf{x}_i), \hat{y}_i)$. From $\hat{y}_i \neq \tilde{y}_i$ we conclude that this agent is a member of C. Therefore, $\hat{y}_i = y_i$ and $\ell(\hat{f}(\mathbf{x}_i), y_i) < \ell(\tilde{f}(\mathbf{x}_i), y_i)$. This contradicts our assumption that no member of C loses from revealing $\tilde{S}$ instead of $\hat{S}$. We emphasize that the proof holds regardless of the values revealed by the agents that are not members of C, and we therefore have group strategyproofness. □

4.2. ERM with other convex loss functions

We have seen that performing ERM with the absolute loss is strategyproof. We now show that the same is not true for most other convex loss functions. Specifically, we examine loss functions of the form $\ell(\alpha, \beta) = \mu(|\alpha - \beta|)$, where $\mu : \mathbb{R}_+ \to \mathbb{R}$ is a monotonically increasing strictly convex function with unbounded subderivatives. Unbounded subderivatives mean that μ cannot be bounded from above by any linear function.

For example, μ can be the function $\mu(\alpha) = \alpha^d$, where d is a real number strictly greater than 1. A popular choice is $d = 2$, which induces the squared loss, $\ell(\alpha, \beta) = (\alpha - \beta)^2$. The following example demonstrates that ERM with the squared loss is not strategyproof.

Example 4.3. Let ℓ be the squared loss function, $\mathcal{X} = \mathbb{R}$, and $\mathcal{F}$ the class of constant functions over $\mathcal{X}$. Further let $S_1 = \{(x_1, 2)\}$ and $S_2 = \{(x_2, 0)\}$. On S, ERM outputs the constant function $\hat{f}(x) \equiv 1$, and agent 1 suffers loss 1. However, if agent 1 reports his value to be 4, ERM outputs $\hat{f}(x) \equiv 2$, with loss of 0 for agent 1.

For every $\mathbf{x} \in \mathcal{X}$, let $\mathcal{F}(\mathbf{x})$ denote the *set of feasible values* at $\mathbf{x}$, formally defined as $\mathcal{F}(\mathbf{x}) = \{f(\mathbf{x}):\ f \in \mathcal{F}\}$. Since $\mathcal{F}$ is a convex set, it follows that $\mathcal{F}(\mathbf{x})$ is either an interval on the real line, a ray, or the entire real line. Similarly, for a multiset $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathcal{X}^n$, denote

$$\mathcal{F}(X) = \left\{\left\langle f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)\right\rangle :\ f \in \mathcal{F}\right\} \subseteq \mathbb{R}^n.$$

We then say that $\mathcal{F}$ is *full* on a multiset $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathcal{X}^n$ if $\mathcal{F}(X) = \mathcal{F}(\mathbf{x}_1) \times \cdots \times \mathcal{F}(\mathbf{x}_n)$. Clearly, requiring that $\mathcal{F}$ is not full on X is a necessary condition for the existence of a training set with points X where one of the agents gains by lying. Otherwise, ERM will fit any set of values for the points with an error of zero. For an example of a function class that is *not* full, consider any function class $\mathcal{F}$ on $\mathcal{X}$, $|\mathcal{F}| \geqslant 2$, and observe that there have to exist $f_1, f_2 \in \mathcal{F}$ and a point $\mathbf{x}_0 \in \mathcal{X}$ such that $f_1(\mathbf{x}_0) \neq f_2(\mathbf{x}_0)$. In this case, $\mathcal{F}$ is not full on any multiset X that contains two copies of $\mathbf{x}_0$.

In addition, if $|\mathcal{F}| = 1$, then any algorithm would trivially be strategyproof irrespective of the loss function. In the following theorem we therefore consider hypothesis classes $\mathcal{F}$ of size at least two which are *not* full on the set X of points of the training set.

Theorem 4.4. *Let $\mu : \mathbb{R}_+ \to \mathbb{R}$ be a monotonically increasing strictly convex function with unbounded subderivatives, and define the loss function $\ell(\alpha, \beta) = \mu(|\alpha - \beta|)$. Let $\mathcal{F}$ be a convex hypothesis class that contains at least two functions, and let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathcal{X}^n$ be a multiset such that $\mathcal{F}$ is not full on X. Then there exist $y_1, \ldots, y_n \in \mathbb{R}$ such that, if $S = \biguplus_{i \in N} S_i$ with $S_i = \{(\mathbf{x}_i, y_i)\}$, ρ_i is degenerate at $\mathbf{x}_i$, and ERM is used, there is an agent who has an incentive to lie.*

An example for a function not covered by this theorem is given by $\nu(\alpha) = \ln(1 + \exp(\alpha))$, which is both monotonic and strictly convex, but has a derivative bounded from above by 1. We use the subderivatives of μ, rather than its derivatives, since we do not require μ to be differentiable.

As before, we actually prove a slightly stronger and more explicit claim about the behavior of the ERM algorithm. The formal proof of Theorem 4.4 follows as a corollary below.

Proposition 4.5. *Let μ and ℓ be as defined in Theorem* 4.4 *and let $\mathcal{F}$ be a convex hypothesis class. Let $\hat{S} = \{(\mathbf{x}_i, \hat{y}_i)\}_{i=1}^m$ be a training set, where $\hat{y}_i \in \mathcal{F}(\mathbf{x}_i)$ for all i, and define $\hat{f} = \mathtt{ERM}(\mathcal{F}, \ell, \hat{S})$. For each $i \in N$, one of the following conditions holds*:

1. $\hat{f}(\mathbf{x}_i) = \hat{y}_i$.
2. *There exists $\tilde{y}_i \in \mathbb{R}$ such that, if we define $\tilde{S} = \hat{S}_{-i} \cup \{(\mathbf{x}_i, \tilde{y}_i)\}$ and $\tilde{f} = \mathtt{ERM}(\mathcal{F}, \ell, \tilde{S})$, $\ell(\tilde{f}(\mathbf{x}_i), \hat{y}_i) < \ell(\hat{f}(\mathbf{x}_i), \hat{y}_i)$.*

To prove the above, we first require a few technical results, which we state in the form of three lemmas. The first lemma takes the perspective of agent i and considers the case where truth-telling results in a function $\hat{f}$ such that $\hat{f}(\mathbf{x}_i) > \hat{y}_i$, i.e., agent i would like the ERM hypothesis to map $\mathbf{x}_i$ to a somewhat lower value. The second lemma then states that there exists a lie that achieves this goal. The gap between the claim of this lemma and the claim of Theorem 4.5 is a subtle one: merely lowering the value of the ERM hypothesis does not necessarily imply a lowering of the loss incurred by agent i. It could be the case that the lie told by agent i caused $\hat{f}(\mathbf{x}_i)$ to become too low, essentially overshooting the desired target value and increasing the loss of agent i. This point is resolved by the third lemma.

Lemma 4.6. *Let ℓ, $\mathcal{F}$, $\hat{S}$ and $\hat{f}$ be as defined in Theorem* 4.5 *and let $i \in N$ be such that $\hat{f}(\mathbf{x}_i) > \hat{y}_i$. Then for all $f \in \mathcal{F}$ for which $f(\mathbf{x}_i) \geqslant \hat{f}(\mathbf{x}_i)$, and for all $y \in \mathbb{R}$ such that $y \leqslant \hat{y}_i$, the dataset $\tilde{S} = \hat{S}_{-i} \cup \{(\mathbf{x}_i, y)\}$ satisfies $\hat{R}(f, \tilde{S}) \geqslant \hat{R}(\hat{f}, \tilde{S})$.*

Proof. Let $f \in \mathcal{F}$ be such that $f(\mathbf{x}_i) \geqslant \hat{f}(\mathbf{x}_i)$, let $y \in \mathbb{R}$ be such that $y \leqslant \hat{y}_i$, and define $\tilde{S} = \hat{S}_{-i} \cup \{(\mathbf{x}_i, y)\}$. We now have that

$$\begin{aligned}
\hat{R}(f, \tilde{S}) &= \hat{R}(f, \tilde{S}_{-i}) + \ell\big(f(\mathbf{x}_i), \tilde{y}_i\big) \\
&= \hat{R}(f, \hat{S}) - \ell\big(f(\mathbf{x}_i), \hat{y}_i\big) + \ell\big(f(\mathbf{x}_i), \tilde{y}_i\big) \\
&= \hat{R}(f, \hat{S}) - \mu\big(f(\mathbf{x}_i) - \hat{y}_i\big) + \mu\big(f(\mathbf{x}_i) - \tilde{y}_i\big). \qquad (11)
\end{aligned}$$

Using the fact that $\hat{f}$ is the empirical risk minimizer with respect to $\hat{S}$, we can get a lower bound for the above and obtain

$$\hat{R}(f, \tilde{S}) \geqslant \hat{R}(\hat{f}, \hat{S}) - \mu\big(f(\mathbf{x}_i) - \hat{y}_i\big) + \mu\big(f(\mathbf{x}_i) - \tilde{y}_i\big).$$

The term $\hat{R}(\hat{f}, \hat{S})$ on the right-hand side can again be rewritten using (11), resulting in

$$\hat{R}(f, \tilde{S}) \geqslant \hat{R}(\hat{f}, \tilde{S}) + \mu\big(\hat{f}(\mathbf{x}_i) - \hat{y}_i\big) - \mu\big(\hat{f}(\mathbf{x}_i) - \tilde{y}_i\big) - \mu\big(f(\mathbf{x}_i) - \hat{y}_i\big) + \mu\big(f(\mathbf{x}_i) - \tilde{y}_i\big).$$

Denoting $a = \hat{f}(\mathbf{x}_i) - \hat{y}_i$, $b = \hat{f}(\mathbf{x}_i) - \tilde{y}_i$, $c = f(\mathbf{x}_i) - \hat{y}_i$, and $d = f(\mathbf{x}_i) - \tilde{y}_i$, we can rewrite the above as

$$\hat{R}(f, \tilde{S}) \geqslant \hat{R}(\hat{f}, \tilde{S}) + \mu(a) - \mu(b) - \mu(c) + \mu(d). \qquad (12)$$

Noting that b, c, and d are all greater than a, and that $b+c-2a = d-a$, we use convexity of μ to obtain

$$\begin{aligned}
\mu(a)+\mu(d) &= \left(\frac{b-a}{d-a}\mu(a)+\frac{c-a}{d-a}\mu(d)\right)+\left(\frac{c-a}{d-a}\mu(a)+\frac{b-a}{d-a}\mu(d)\right)\\
&\geqslant \mu\left(\frac{(b-a)a+(c-a)d}{d-a}\right)+\mu\left(\frac{(c-a)a+(b-a)d}{d-a}\right)\\
&= \mu\left(\frac{(b+c-2a)a+(c-a)(d-a)}{d-a}\right)+\mu\left(\frac{(c+b-2a)a+(b-a)(d-a)}{d-a}\right)\\
&= \mu(c)+\mu(b).
\end{aligned}$$

Combining this inequality with (12) concludes the proof. □

Lemma 4.7. *Let ℓ, $\mathcal{F}$, $\hat{S}$ and $\hat{f}$ be as defined in Theorem 4.5 and let $i \in N$ be such that $\hat{f}(\mathbf{x}_i) > \hat{y}_i$. Then there exists $\tilde{y}_i \in \mathbb{R}$ such that if we define $\tilde{S} = \hat{S}_{-i} \cup \{(\mathbf{x}_i, \tilde{y}_i)\}$ and $\tilde{f} = \mathtt{ERM}(\mathcal{F}, \ell, \tilde{S})$, then $\tilde{f}(\mathbf{x}_i) < \hat{f}(\mathbf{x}_i)$.*

Proof. Let i be such that $\hat{f}(\mathbf{x}_i) \neq \hat{y}_i$ and assume without loss of generality that $\hat{f}(\mathbf{x}_i) > \hat{y}_i$. Since $\hat{y}_i \in \mathcal{F}(\mathbf{x}_i)$, there exists a function $f' \in \mathcal{F}$ such that $f'(\mathbf{x}_i) = \hat{y}_i$. Now define

$$\phi = \frac{\hat{R}(f', \hat{S}_{-i}) - \hat{R}(\hat{f}, \hat{S}_{-i}) + 1}{\hat{f}(\mathbf{x}_i) - f'(\mathbf{x}_i)}. \tag{13}$$

It holds, by definition, that $\hat{R}(f', \hat{S}) > \hat{R}(\hat{f}, \hat{S})$ and that $\ell(f'(\mathbf{x}_i), \hat{y}_i) < \ell(\hat{f}(\mathbf{x}_i), \hat{y}_i)$, and therefore the numerator of (13) is positive. Furthermore, our assumption implies that the denominator of (13) is also positive, so ϕ is positive as well.

Since μ has unbounded subderivatives, there exists $\psi > 0$ large enough such that the subderivative of μ at ψ is greater than ϕ. By the definition of the subderivative, we have that

$$\text{for all } \alpha \geqslant \psi, \quad \mu(\psi) + (\alpha - \psi)\phi \leqslant \mu(\alpha). \tag{14}$$

Defining $\tilde{y}_i = f'(\mathbf{x}_i) - \psi$ and $\tilde{S} = \hat{S}_{-i} \cup \{(\mathbf{x}_i, \tilde{y}_i)\}$, we have that

$$\ell\big(f'(\mathbf{x}_i), \tilde{y}_i\big) = \mu\big(f'(\mathbf{x}_i) - \tilde{y}_i\big) = \mu(\psi),$$

and therefore

$$\hat{R}\big(f', \tilde{S}\big) = \hat{R}\big(f', \tilde{S}_{-i}\big) + \ell\big(f'(\mathbf{x}_i), \tilde{y}_i\big) = \hat{R}\big(f', \tilde{S}_{-i}\big) + \mu(\psi). \tag{15}$$

We further have that

$$\ell\big(\hat{f}(\mathbf{x}_i), \tilde{y}_i\big) = \mu\big(\hat{f}(\mathbf{x}_i) - \tilde{y}_i\big) = \mu\big(\hat{f}(\mathbf{x}_i) - f'(\mathbf{x}_i) + \psi\big).$$

Combining (14) with the fact that $\hat{f}(\mathbf{x}_i) - f'(\mathbf{x}_i) > 0$, we get $\mu(\psi) + (\hat{f}(\mathbf{x}_i) - f'(\mathbf{x}_i))\phi$ as a lower bound for the above. Plugging in the definition of ϕ from (13), we obtain

$$\ell\big(\hat{f}(\mathbf{x}_i), \tilde{y}_i\big) \geqslant \mu(\psi) + \hat{R}\big(f', \hat{S}_{-i}\big) - \hat{R}(\hat{f}, \hat{S}_{-i}) + 1,$$

and therefore,

$$\hat{R}(\hat{f}, \tilde{S}) = \hat{R}(\hat{f}, \tilde{S}_{-i}) + \ell\big(\hat{f}(\mathbf{x}_i), \tilde{y}_i\big) \geqslant \mu(\psi) + \hat{R}\big(f', \hat{S}_{-i}\big) + 1.$$

Comparing the above with (15), we get

$$\hat{R}(\hat{f}, \tilde{S}) > \hat{R}\big(f', \tilde{S}\big).$$

We now use Lemma 4.6 to extend the above to every $f \in \mathcal{F}$ for which $f(\mathbf{x}_i) \geqslant \hat{f}(\mathbf{x}_i)$, namely, we now have that any such f satisfies $\hat{R}(f, \tilde{S}) > \hat{R}(f', \tilde{S})$. We conclude that the empirical risk minimizer $\tilde{f}$ must satisfy $\tilde{f}(\mathbf{x}_i) < \hat{f}(\mathbf{x}_i)$. □

Lemma 4.8. *Let ℓ and $\mathcal{F}$ be as defined in Theorem 4.5, let $S = \{(\mathbf{x}_i, \hat{y}_i)\}_{i=1}^m$ be a dataset, and let $i \in N$ be an arbitrary index. Then the function $g(\tilde{y}) = f(\mathbf{x}_i)$, where $f = \mathtt{ERM}(\mathcal{F}, \ell, S_{-i} \cup \{(\mathbf{x}_i, \tilde{y})\})$, is continuous.*

Proof. We first restate ERM as a minimization problem over vectors in $\mathbb{R}^m$. Define the set of feasible values for the points $\mathbf{x}_1, \ldots, \mathbf{x}_m$ to be

$$\mathcal{G} = \big\{\big(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_m)\big)\colon\ f \in \mathcal{F}\big\}.$$

Our assumption that $\mathcal{F}$ is a convex set implies that $\mathcal{G}$ is a convex set as well. Now, define the function

$$L(\mathbf{v}, \tilde{y}) = \ell(v_i, \tilde{y}) + \sum_{j \neq i} \ell(v_j, \hat{y}_j), \quad \text{where } \mathbf{v} = (v_1, \ldots, v_m).$$

Finding $f \in \mathcal{F}$ that minimizes the empirical risk with respect to the dataset $S_{-i} \cup \{(\mathbf{x}_i, \tilde{y})\}$ is equivalent to calculating $\min_{\mathbf{v} \in \mathcal{G}} L(\mathbf{v}, \tilde{y})$. Moreover, $g(\tilde{y})$ can be equivalently defined as the value of the i'th coordinate of the vector in $\mathcal{G}$ that minimizes $L(\mathbf{v}, \tilde{y})$.

To prove that g is continuous at an arbitrary point $\tilde{y} \in \mathbb{R}$, we show that for every $\epsilon > 0$ there exists $\delta > 0$ such that if $y \in [\tilde{y} - \delta, \tilde{y} + \delta]$ then $g(y) \in [g(\tilde{y}) - \epsilon, g(\tilde{y}) + \epsilon]$. For this, let $\tilde{y}$ and $\epsilon > 0$ be arbitrary real numbers, and define

$$\mathbf{u} = \arg\min_{\mathbf{v} \in \mathcal{G}} L(\mathbf{v}, \tilde{y}).$$

Since ℓ is strictly convex in its first argument, so is L. Consequently, $\mathbf{u}$ is the unique global minimizer of L. Also define

$$\mathcal{G}_\epsilon = \big\{\mathbf{v} \in \mathcal{G}\colon\ |v_i - u_i| \geqslant \epsilon\big\}.$$

Assume that ϵ is small enough that $\mathcal{G}_\epsilon$ is not empty (if no such ϵ exists, the lemma holds trivially). Note that $\mathbf{u} \notin \mathcal{G}_\epsilon$ for any value of $\epsilon > 0$. Define $\bar{\mathcal{G}}_\epsilon$ to be the closure of $\mathcal{G}_\epsilon$ and let

$$\nu = \inf_{\mathbf{v} \in \bar{\mathcal{G}}_\epsilon} L(\mathbf{v}, \tilde{y}) - L(\mathbf{u}, \tilde{y}).$$

Since μ is strictly convex and has unbounded subderivatives, the level-sets of $L(\mathbf{v}, \tilde{y})$, as a function of $\mathbf{v}$, are all bounded. Therefore, there exists $\mathbf{w} \in \bar{\mathcal{G}}_\epsilon$ that attains the infimum above. More precisely, $\mathbf{w}$ is such that $L(\mathbf{w}, \tilde{y}) - L(\mathbf{u}, \tilde{y}) = \nu$. Using uniqueness of the minimizer $\mathbf{u}$, as well as the fact that $\mathbf{w} \neq \mathbf{u}$, we conclude that $\nu > 0$. We have proven that if $\mathbf{v} \in \mathcal{F}$ is such that

$$L(\mathbf{v}, \tilde{y}) < L(\mathbf{u}, \tilde{y}) + \nu, \tag{16}$$

then $|v_i - u_i| < \epsilon$. It therefore suffices to show that there exists $\delta > 0$ such that if $y \in [\tilde{y} - \delta, \tilde{y} + \delta]$ then the vector $\mathbf{v} \in \mathcal{G}$ that minimizes $L(\mathbf{v}, y)$ satisfies the condition in (16).

Since ℓ is convex in its second argument, ℓ is also continuous in its second argument. Thus, there exists $\delta > 0$ such that for all $y \in [\tilde{y} - \delta, \tilde{y} + \delta]$ it holds that both

$$\ell(u_i, \tilde{y}) < \ell(u_i, y) + \nu/2 \quad \text{and} \quad \ell(w_i, y) < \ell(w_i, \tilde{y}) + \nu/2,$$

where $\mathbf{w} = \arg\min_{\mathbf{v} \in \mathcal{G}} L(\mathbf{v}, y)$. Therefore,

$$L(\mathbf{u}, \tilde{y}) < L(\mathbf{u}, y) + \nu/2 \quad \text{and} \quad L(\mathbf{w}, \tilde{y}) < L(\mathbf{w}, y) + \nu/2.$$

Finally, since $\mathbf{w}$ minimizes $L(\mathbf{v}, y)$, we have $L(\mathbf{w}, y) \leqslant L(\mathbf{u}, y)$. Combining these three inequalities yields the condition in (16). □

We are now ready to prove Proposition 4.5, and then Theorem 4.4.

Proof of Proposition 4.5. If $\hat{f}(\mathbf{x}_i) = \hat{y}_i$ for all $i \in N$, we are done. Otherwise let i be an index for which $\hat{f}(\mathbf{x}_i) \neq \hat{y}_i$ and assume without loss of generality that $\hat{f}(\mathbf{x}_i) > \hat{y}_i$. Using Lemma 4.7, we know that there exists $\tilde{y}_i \in \mathbb{R}$ such that if we define $\tilde{S} = \hat{S}_{-i} \cup \{(\mathbf{x}_i, \tilde{y}_i)\}$ and $f' = \mathtt{ERM}(\mathcal{F}, \ell, \tilde{S})$, then f' satisfies $\hat{f}(\mathbf{x}_i) > f'(\mathbf{x}_i)$.

We consider the two possible cases: either $\hat{f}(\mathbf{x}_i) > f'(\mathbf{x}_i) \geqslant \hat{y}_i$, and therefore $\ell(\hat{f}(\mathbf{x}_i), \hat{y}_i) > \ell(f'(\mathbf{x}_i), \hat{y}_i)$ as required. Otherwise, $\hat{f}(\mathbf{x}_i) > \hat{y}_i > f'(\mathbf{x}_i)$. Using Lemma 4.8, we know that $f(\mathbf{x}_i)$ changes continuously with $\tilde{y}_i$, where $f = \mathtt{ERM}(\mathcal{F}, \ell, S_{-i} \cup \{(\mathbf{x}_i, \tilde{y}_i)\})$. Relying on the elementary *Intermediate Value Theorem*, we conclude that for some $y \in [\hat{y}_i, \tilde{y}_i]$ it holds that f, the empirical risk minimizer with respect to the dataset $S_{-i} \cup \{(\mathbf{x}_i, y)\}$, satisfies $f(\mathbf{x}_i) = \hat{y}_i$. Once again we have $\ell(\hat{f}(\mathbf{x}_i), \hat{y}_i) > \ell(f(\mathbf{x}_i), \hat{y}_i)$. □

Proof of Theorem 4.4. Since $\mathcal{F}$ is not full on X, there are $y_1^*, \ldots, y_n^*$ such that $y_i^* \in \mathcal{F}(\mathbf{x}_i)$ for all i, and $\langle y_1^*, \ldots, y_n^* \rangle \notin \mathcal{F}(X)$. Defining $S = \{(\mathbf{x}_i, y_i^*)\}_{i=1}^n$, there exists some agent i which isn't satisfied by the output of the ERM algorithm on S. Using Proposition 4.5 we conclude that this agent has an incentive to lie. □

It is natural to ask what happens for loss functions that are *sublinear* in the sense that they cannot be bounded from below by any linear function with strictly positive derivative. A property of such loss functions, and the reason why they are rarely used in practice, is that the set of empirical risk minimizers need no longer be convex. It is thus unclear how tie-breaking should be defined in order to find a unique empirical risk minimizer. Furthermore, the following example provides a negative answer to the question of general strategyproofness of ERM with sublinear loss.

Example 4.9. We demonstrate that ERM is not strategyproof if $\ell(a,b) = \sqrt{|a-b|}$ and $\mathcal{F}$ is the class of constant functions over $\mathbb{R}$. Let $S = \{(x_1,1),(x_2,2),(x_3,4),(x_4,6)\}$ and $\tilde{S} = \{(x_1,1),(x_2,2),(x_3,4),(x_4,4)\}$. Clearly, the local minima of $\hat{R}(f,S)$ and $\hat{R}(f,\tilde{S})$ have the form $f(x) \equiv y$ where $(x_i,y) \in S$ or $(x_i,y) \in \tilde{S}$, respectively, for some $i \in \{1,2,3,4\}$. The empirical risk minimizer for S is the constant function $f_1(x) \equiv 2$, while that for $\tilde{S}$ is $f_2(x) \equiv 4$. Thus, agent 4 can declare his value to be 4 instead of 6 to decrease his loss from 2 to $\sqrt{2}$.

5. Uniform distributions over the sample

We now turn to settings where a single agent holds a (possibly) non-degenerate distribution over the input space. However, we still do not move to the full level of generality. Rather, we concentrate on a setting where for each agent i, ρ_i is the uniform distribution over the sample points x_{ij}, $j = 1,\ldots,m$. While this setting is equivalent to curve fitting with multiple agents and may be interesting in its own right, we primarily engage in this sort of analysis as a stepping stone in our quest to understand the learning game. The results in this section will function as building blocks for the results of Section 6.

Since each agent $i \in N$ now holds a uniform distribution over his sample, we can simply assume that his cost is his average empirical loss on the sample, $\hat{R}(\tilde{f},S_i) = (1/m)\cdot\sum_{j=1}^{m}\ell(\tilde{f}(\mathbf{x}_{ij}),y_{ij})$. The mechanism's goal is to minimize $\hat{R}(\tilde{f},S)$. We stress at this point that the results in this section also hold if the agents' samples differ in size. This is of course true for the negative results, but also holds for the positive ones. As we move to this more general setting, truthfulness of ERM immediately becomes a thorny issue even under absolute loss. Indeed, the next example indicates that more sophisticated mechanisms must be used to achieve strategyproofness.

Example 5.1. Let $\mathcal{F}$ be the class of constant functions over $\mathbb{R}^k$, $N = \{1,2\}$, and assume the absolute loss function is used. Let $S_1 = \{(1,1),(2,1),(3,0)\}$ and $S_2 = \{(4,0),(5,0),(6,1)\}$. The global empirical risk minimizer (according to our tie-breaking rule) is the constant function $f_1(x) \equiv 0$ with $\hat{R}(f_1,S_1) = 2/3$. However, if agent 1 declares $\tilde{S}_1 = \{(1,1),(2,1),(3,1)\}$, then the empirical risk minimizer becomes $f_2(x) \equiv 1$, which is the optimal fit for agent 1 with $\hat{R}(f_2,S_1) = 1/3$.

5.1. Mechanisms with payments

One possibility to overcome the issue that became manifest in Example 5.1 is to consider mechanisms that not only return an allocation, but can also transfer payments to and from the agents based on the inputs they provide. A famous example for such a payment rule is the Vickrey–Clarke–Groves (VCG) mechanism [16–18]. This mechanism, which has been described in detail in Section 2, starts from an efficient allocation and computes each agent's payment according to the utility of the other agents, thus aligning the individual interests of each agent with that of society.

In our setting, where social welfare equals the total empirical risk, ERM generates a function, or outcome, that maximizes social welfare and can therefore be directly augmented with VCG payments. Given an outcome $\hat{f}$, each agent i has to pay an amount of $\hat{R}(\hat{f},\tilde{S}_{-i})$. In turn, the agent can receive some amount $h_i(\tilde{S}_{-i})$ that does *not* depend on the values he has reported, but possibly on the values reported by the other agents. It is well known [18], and also easily verified, that this family of mechanisms is strategyproof: no agent is motivated to lie regardless of the other agents' actions. Furthermore, this result holds for any loss function, and may thus be an excellent solution for some settings.

In many other settings, however, especially in the world of the Internet, transferring payments to and from users can pose serious problems, up to the extent that it might become completely infeasible. The practicality of VCG payments in particular has recently also been disputed for various other reasons [19]. Perhaps most relevant to our work is the fact that VCG mechanisms are in general susceptible to manipulation by coalitions of agents and thus not group strategyproof. It is therefore worthwhile to explore which results can be obtained when payments are disallowed. This will be the subject of the following section.

5.2. Mechanisms without payments

In this section, we restrict ourselves to the absolute loss function. When ERM is used, and for the special case covered in Section 4, this function was shown to possess incentive properties far superior to any other loss function. This fuels hope that similar incentive compatibility results can be obtained with uniform distributions over the samples, even when payments are disallowed. This does not necessarily mean that good mechanisms without payments cannot be designed for other loss functions, even in the more general setting of this section. We leave the study of such mechanisms for future work.

ERM is *efficient*, i.e., it minimizes the overall loss and maximizes social welfare. In light of Example 5.1, we shall now sacrifice efficiency for incentive compatibility. More precisely, we seek strategyproof or group strategyproof mechanisms which are at the same time *approximately efficient*. We should stress that the reason we resort to approximation is not to make the mechanism computationally tractable, but to achieve incentive compatibility without payments, like we had in Section 4.

Example 5.1, despite its simplicity, is surprisingly robust against many conceivably truthful mechanisms. The reader may have noticed, however, that the values of the agents in this example are not "individually realizable": in particular, there is no constant function which *realizes* agent 1's values, i.e., fits them with a loss of zero. In fact, agent 1 benefits from revealing values which are consistent with his individual empirical risk minimizer. This insight leads us to design the following simple but useful mechanism, which we will term "project-and-fit":

Input: A hypothesis class $\mathcal{F}$ and a sample $S = \biguplus S_i$, $S_i \subseteq \mathcal{X} \times \mathbb{R}$.
Output: A function $f \in \mathcal{F}$.
Mechanism:

1. For each $i \in N$, let $f_i = \texttt{ERM}(\mathcal{F}, S_i)$.
2. Define $\tilde{S}_i = \{(\mathbf{x}_{i1}, f_i(x_{i1})), \ldots, (\mathbf{x}_{im}, f_i(x_{im}))\}$.
3. Return $f = \texttt{ERM}(\tilde{S})$, where $\tilde{S} = \biguplus_{i=1}^{n} \tilde{S}_i$.

In other words, the mechanism calculates the individual empirical risk minimizer for each agent and uses it to relabel the agent's sample. Then, the relabeled samples are combined, and ERM is performed. It is immediately evident that this mechanism achieves group strategyproofness at least with respect to Example 5.1.

More generally, it can be shown that the mechanism is group strategyproof when $\mathcal{F}$ is the class of constant functions over $\mathbb{R}^k$. Indeed, it is natural to view our setting through the eyes of social choice theory (see, e.g., [20]): agents entertain (weak) preferences over a set of alternatives, i.e., the functions in $\mathcal{F}$. In the case of constant functions, agents' preferences are what is known as *single-plateau* [21]: each agent has an interval of ideal points minimizing his individual empirical risk, and moving away from this plateau in either direction strictly decreases the agent's utility. More formally, let a_1, a_2 be constants such that the constant function $f(x) \equiv a$ minimizes an agent's empirical risk if and only if $a \in [a_1, a_2]$. If a_3 and a_4 satisfy $a_3 < a_4 \leqslant a_1$ or $a_3 > a_4 \geqslant a_2$, then the agent strictly prefers the constant function a_4 to the constant function a_3. As such, single-plateau preferences generalize the class of single-peaked preferences. For dealing with single-plateau preferences, Moulin [21] defines the class of generalized Condorcet winner choice functions, and shows that these are group strategyproof.

When $\mathcal{F}$ is the class of constant functions and ℓ is the absolute loss, the constant function equal to a median value in a sample S minimizes the empirical risk with respect to S. This is because there must be at least as many values below the median value as are above, and thus moving the fit upward (or downward) must monotonically increase the sum of distances to the values. Via tie-breaking, project-and-fit essentially turns the single-plateau preferences into single-peaked ones, and then chooses the median peak. Once again, group strategyproofness follows from the fact that an agent can only change the mechanism's output by increasing its distance from his own empirical risk minimizer.

Quite surprisingly, project-and-fit is not only truthful but also provides a constant approximation ratio when $\mathcal{F}$ is the class of constant functions or the class of homogeneous linear functions over $\mathbb{R}$, i.e., functions of the form $f(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x}$. The class of homogeneous linear functions, in particular, is important in machine learning, for instance in the context of Support Vector Machines [22].

Theorem 5.2. *Assume that $\mathcal{F}$ is the class of constant functions over $\mathbb{R}^k$, $k \in \mathbb{N}$ or the class of homogeneous linear functions over $\mathbb{R}$. Then project-and-fit is group strategyproof and 3-efficient.*

Proof. We shall first prove the theorem for the case when $\mathcal{F}$ is the class of constant functions over $\mathbb{R}^k$ (Steps 1 and 2), and then extend the result to homogeneous linear functions over $\mathbb{R}$ (Step 3). We have already shown truthfulness, and therefore directly turn to approximate efficiency. In the following, we denote the empirical risk minimizer by $f^*(x) \equiv a^*$, and the function returned by project-and-fit by $f(x) \equiv a$.

Step 1: $|\{y_{ij}\colon\ y_{ij} \leqslant a\}| \geqslant \frac{1}{4}nm$ and $|\{y_{ij}\colon\ y_{ij} \geqslant a\}| \geqslant \frac{1}{4}nm$. Let $\tilde{y}_{ij}$ denote the projected values of agent i. As noted above, when $\mathcal{F}$ is the class of constant functions, the mechanism in fact returns the median of the values $\tilde{y}_{ij}$, and thus

$$\left|\{\tilde{y}_{ij}\colon\ \tilde{y}_{ij} \leqslant a\}\right| \geqslant \frac{1}{2}nm. \tag{17}$$

Furthermore, since for all j, $\tilde{y}_{ij}$ is the median of the original values y_{ij} of agent i, it must hold that at least half of these values are smaller than their corresponding original value, i.e.,

$$\left|\{y_{ij}\colon\ y_{ij} \leqslant a\}\right| \geqslant \frac{1}{2}\left|\{\tilde{y}_{ij}\colon\ \tilde{y}_{ij} \leqslant a\}\right|. \tag{18}$$

Combining (17) and (18), we obtain $|\{y_{ij}\colon\ y_{ij} \leqslant a\}| \geqslant \frac{1}{4}nm$. By symmetrical arguments, we get that $|\{y_{ij}\colon\ y_{ij} \geqslant a\}| \geqslant \frac{1}{4}nm$.

Step 2: 3-efficiency for constant functions. Denote $d = |a - a^*|$, and assume without loss of generality that $a < a^*$. We now have that

$$\begin{aligned}
\hat{R}(f, S) &= \frac{1}{nm} \sum_{i,j} |y_{ij} - a| \\
&= \frac{1}{nm} \Bigg(\sum_{i,j:\ y_{ij} \leqslant a} (a - y_{ij}) + \sum_{i,j:\ a < y_{ij} \leqslant a^*} (y_{ij} - a) + \sum_{i,j:\ y_{ij} > a^*} (y_{ij} - a) \Bigg) \\
&\leqslant \frac{1}{nm} \Bigg(\sum_{i,j:\ y_{ij} \leqslant a} (a - y_{ij}) + \sum_{i,j:\ a < y_{ij} \leqslant a^*} d + \sum_{i,j:\ y_{ij} > a^*} \big(d + (y_{ij} - a^*)\big) \Bigg) \\
&= \frac{1}{nm} \Bigg(\sum_{i,j:\ y_{ij} \leqslant a} (a - y_{ij}) + \sum_{i,j:\ y_{ij} > a^*} (y_{ij} - a^*) + \big|\{i, j:\ y_{ij} > a\}\big| \cdot d \Bigg).
\end{aligned}$$

We now bound the last expression above by replacing $|\{i, j:\ y_{ij} > a\}|$ with its upper bound $\frac{3}{4}nm$ derived in Step 1 and obtain

$$\hat{R}(f, S) \leqslant \frac{1}{nm} \Bigg(\sum_{i,j:\ y_{ij} \leqslant a} (a - y_{ij}) + \sum_{i,j:\ y_{ij} > a^*} (y_{ij} - a^*) + \frac{3}{4} nm \cdot d \Bigg).$$

Similarly,

$$\hat{R}(f^*, S) \geqslant \frac{1}{nm} \Bigg(\sum_{i,j:\ y_{ij} \leqslant a} \big(d + (a - y_{ij})\big) + \sum_{i,j:\ y_{ij} > a^*} (y_{ij} - a^*) \Bigg),$$

and using Step 1,

$$\hat{R}(f^*, S) \geqslant \frac{1}{nm} \Bigg(\sum_{i,j:\ y_{ij} \leqslant a} (a - y_{ij}) + \sum_{i,j:\ y_{ij} > a^*} (y_{ij} - a^*) + \frac{1}{4} nm \cdot d \Bigg).$$

Since two of the expressions in the upper bound for $\hat{R}(f, S)$ and the lower bound for $\hat{R}(f^*, S)$ are identical, it is now self-evident that $\hat{R}(f, S)/\hat{R}(f^*, S) \leqslant 3$.

Step 3: Extension to homogeneous linear functions over $\mathbb{R}$. We describe a reduction from the case of homogeneous functions over $\mathbb{R}$ to the case of constant functions over $\mathbb{R}$. Given a sample S, we create a sample S' by mapping each example $(x, y) \in S$ to $|x|$ copies of the example $(x, y/x)$.[4] Let f_1 be the homogeneous linear function defined by $f_1(x) = a \cdot x$, and let f_2 be the constant function defined by $f_2(x) = a$. It is now straightforward to show that $\hat{R}(f_1, S) = \hat{R}(f_2, S')$, and that project-and-fit chooses f_1 when given the class of homogeneous linear functions and S if and only if it chooses f_2 when given the class of constant functions and S'. □

A simple example shows that the 3-efficiency analysis given in the proof is tight. We generalize this observation by proving that, for the class of constant or homogeneous linear functions and irrespective of the dimension of $\mathcal{X}$, no truthful mechanism without payments can achieve an efficiency ratio better than 3. It should be noted that this lower bound holds for any choice of points x_{ij}.

Theorem 5.3. *Let $\mathcal{F}$ be the class of constant functions over $\mathbb{R}^k$ or the class of homogeneous linear functions over $\mathbb{R}^k$, $k \in \mathbb{N}$. Then there exists* no *strategyproof mechanism without payments that is $(3 - \epsilon)$-efficient for any $\epsilon > 0$, even when $|N| = 2$.*

We first require a technical result. For this, assume that $\mathcal{F}$ is the class of constant functions over $\mathbb{R}^k$, let $N = \{1, 2\}$, and fix some truthful mechanism M.

Lemma 5.4. *Let $q, t \in \mathbb{N}$, and define $m = 2t - 1$. Then there exists a sample S defined by*

$$S_1 = \big\{(\mathbf{x}_{11}, y), (\mathbf{x}_{12}, y), \ldots, (\mathbf{x}_{1m}, y)\big\} \quad \text{and} \quad S_2 = \big\{\big(\mathbf{x}_{21}, y'\big), \big(\mathbf{x}_{22}, y'\big), \ldots, \big(\mathbf{x}_{2m}, y'\big)\big\},$$

such that $y - y' = 2^q$ and $M(S) \geqslant y - \frac{1}{2}$ or $M(S) \leqslant y' + \frac{1}{2}$.

[4] Here we assume that the values x are integers, but it is possible to deal with noninteger values by assigning weights.

Proof. We perform an induction on q. For $q = 0$, we simply set $y = 1$ and $y' = 0$. Now, let S be a sample as in the formulation of the lemma, and let $a = M(S)$, i.e., a is the constant function returned by M given S. We distinguish two different cases.

Case 1: If $a \geqslant y - 1/2$, let S' such that $S'_1 = S_1$ and

$$S'_2 = \big\{(\mathbf{x}_{21}, 2y' - y), \ldots, (\mathbf{x}_{2m}, 2y' - y)\big\}.$$

Notice that $y - (2y' - y) = 2(y - y')$, so the distance between the values has doubled. Denote $a' = M(S')$. Due to truthfulness of M, it must hold that $\ell(a', y') \geqslant \ell(a, y') \geqslant 2^q - \frac{1}{2}$. Otherwise, if agent 2's true type was S_2, he would benefit by saying that his type is in fact S'_2. Therefore, $a' \geqslant y - \frac{1}{2}$ or $a' \leqslant y' - (2^q + \frac{1}{2}) = 2y' - y + \frac{1}{2}$.

Case 2: If $a \leqslant y' + \frac{1}{2}$, let S' such that $S'_2 = S_2$ and

$$S'_1 = \big\{(\mathbf{x}_{11}, 2y - y'), \ldots, (\mathbf{x}_{1m}, 2y - y')\big\}.$$

Analogously to Case 1, the induction step follows from truthfulness of M with respect to agent 1. □

Proof of Theorem 5.3. Consider the sample S as in the statement of the lemma, and assume without loss of generality that $M(S) = a \geqslant y - \frac{1}{2}$. Otherwise, symmetrical arguments apply. We first observe that if M is approximately efficient, it cannot be the case that $M(S) > y$. Otherwise, let S' be the sample such that $S'_1 = S_1$ and

$$S'_2 = \big\{(\mathbf{x}_{21}, y), \ldots, (\mathbf{x}_{2m}, y)\big\},$$

and denote $a' = M(S')$. Then, by truthfulness with respect to agent 2, $\ell(a', y') \geqslant \ell(a, y')$. It follows that $a' \neq y$, and therefore $\hat{R}(a', S') > 0$. Since $\hat{R}(y, S') = 0$, the efficiency ratio is not bounded.

Now let S'' be such that $S''_2 = S_2$, and

$$S''_1 = \big\{(\mathbf{x}_{11}, y), (\mathbf{x}_{12}, y), \ldots, (\mathbf{x}_{1t}, y), (\mathbf{x}_{1,t+1}, y'), \ldots, (\mathbf{x}_{1m}, y')\big\},$$

i.e., agent 1 has t points at y and $t - 1$ points at y'. Let $a'' = M(S'')$. Due to truthfulness, it must hold that $\ell(a'', y) = \ell(a, y)$, since agent 1's empirical risk minimizer with respect to both S and S'' is y. Since we already know that $y - \frac{1}{2} \leqslant a \leqslant y$, we get that $a'' \geqslant y - \frac{1}{2}$, and thus $\hat{R}(a'', S'') \geqslant \frac{(3t-2)}{(4t-2)}(2^q - \frac{1}{2})$. On the other hand, the empirical risk minimizer on S'' is y', and $\hat{R}(y', S'') \leqslant \frac{t}{4t-2}2^q$. The efficiency ratio $\hat{R}(a'', S'')/\hat{R}(y', S'')$ tends to 3 as t and q tend to infinity.

We will now explain how this result can be extended to homogeneous linear functions over $\mathbb{R}^k$. For this, define the sample S by

$$S_1 = \big\{(\langle t - 1, 0, \ldots, 0\rangle, (t-1)y), (\langle t, 0, \ldots, 0\rangle, ty)\big\} \quad \text{and}$$
$$S_2 = \big\{(\langle t - 1, 0, \ldots, 0\rangle, (t-1)y'), (\langle t, 0, \ldots, 0\rangle, ty')\big\}.$$

As with constant functions, a homogeneous linear function defined by $\mathbf{a}$ satisfies $\hat{R}(\mathbf{a}, S_1) = |a_1 - y|$, and $\hat{R}(\mathbf{a}, S_2) = |a_1 - y'|$. Therefore, we can use similar arguments to the ones above to show that there exists a sample S with $y - y' = 2^q$, and if $\mathbf{a} = M(S)$ for some truthful mechanism M, then $y - \frac{1}{2} \leqslant a_1 \leqslant y$ or $y' \leqslant a_1 \leqslant y' + \frac{1}{2}$. As before, we complete the proof by splitting the points controlled by agent 1, i.e., by considering the sample S' where $S'_1 = \{(\langle t - 1, 0, \ldots, 0\rangle, (t-1)y'), (\langle t, 0, \ldots, 0\rangle, ty)\}$. □

Let us recapitulate. We have found a group strategyproof and 3-efficient mechanism for the class of constant functions over $\mathbb{R}^k$ and for the class of homogeneous linear functions over $\mathbb{R}$. A matching lower bound, which also applies to multi-dimensional homogeneous linear functions, shows that this result cannot be improved upon for these classes. It is natural to ask at this point if project-and-fit remains strategyproof when considering more complex hypothesis classes, such as homogeneous linear functions over $\mathbb{R}^k$, $k \geqslant 2$, or linear functions. An example serves to answer this question in the negative.

Example 5.5. We demonstrate that project-and-fit is not strategyproof when $\mathcal{F}$ is the class of linear functions over $\mathbb{R}$. Let $S_1 = \{(0, 0), (4, 1)\}$ and $S_2 = \{(1, 1), (2, 0)\}$. Since S_1 and S_2 are individually realizable, the mechanism simply returns the empirical risk minimizer, which is $f(x) = x/4$ (this can be determined by solving a linear program). It further holds that $\hat{R}(f, S_2) = 5/8$. If, however, one considers $\tilde{S}_2 = \{(1, 1), (2, 1)\}$ and the same S_1, then the mechanism returns $\tilde{f}(x) = 1$. Agent 2 benefits from this lie as $\hat{R}(\tilde{f}, S_2) = 1/2$.

It is also possible to extend this example to the case of homogeneous linear functions over $\mathbb{R}^2$ by fixing the second coordinate of all points at 1, i.e., mapping each $x \in \mathbb{R}$ to $\mathbf{x}' = (x, 1) \in \mathbb{R}^2$. Indeed, the value of a homogeneous linear function $f(\mathbf{x}) = \langle a, b\rangle \cdot \mathbf{x}$ on the point $(x, 1)$ is $ax + b$.

Is there some other mechanism which deals with more complex hypothesis classes and provides a truthful approximation? In order to tackle this question, it will be instructive to once again view the hypothesis class $\mathcal{F}$ as a set of alternatives. The agents' types induce a preference order over this set of alternatives. Explicitly, agent i weakly prefers function f_1 to function f_2 if and only if $\hat{R}(f_1, S_i) \leqslant \hat{R}(f_2, S_i)$. A mechanism without payments is a social choice function from the agents' preferences over $\mathcal{F}$ to $\mathcal{F}$.

The celebrated Gibbard–Satterthwaite Theorem [23,24] asserts that every truthful social choice function from the set of *all* linear preferences over some set A of alternatives to A must be *dictatorial*, in the sense that there is some agent d such that the social outcome is always the one most preferred by d. Observe that this theorem does not directly apply in our case, since voters' preferences are restricted to a strict subset of all possible preference relations over $\mathcal{F}$.

For the time being, let us focus on homogeneous linear functions f over $\mathbb{R}^k$, $k \geqslant 2$. This class is isomorphic to $\mathbb{R}^k$, as every such function can be represented by a vector $\mathbf{a} \in \mathbb{R}^k$ such that $f(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x}$. Let R be a weak preference relation over $\mathbb{R}^k$, and let P be the asymmetric part of R (i.e., $\mathbf{a}P\mathbf{a}'$ if and only if $\mathbf{a}R\mathbf{a}'$ and not $\mathbf{a}'R\mathbf{a}$). R is called *star-shaped* if there is a unique point $\mathbf{a}^* \in \mathbb{R}^k$ such that for all $\mathbf{a} \in \mathbb{R}^k$ and $\lambda \in (0,1)$, $\mathbf{a}^* P(\lambda\mathbf{a}^* + (1-\lambda)\mathbf{a})P\mathbf{a}$. In our case preferences are clearly star-shaped, as for any $\mathbf{a}, \mathbf{a}' \in \mathbb{R}^k$ and any sample S, $\hat{R}((\lambda\mathbf{a} + (1-\lambda)\mathbf{a}'), S) = \lambda\hat{R}(\mathbf{a}, S) + (1-\lambda)\hat{R}(\mathbf{a}', S)$.

A preference relation R over $\mathbb{R}^m$ is called *separable* if for every j, $1 \leqslant j \leqslant m$, all $x, y \in \mathbb{R}^m$, and all $a_j, b_j \in \mathbb{R}$,

$$\langle x_{-j}, a_j\rangle R \langle x_{-j}, b_j\rangle \quad \text{if and only if} \quad \langle y_{-j}, a_j\rangle R \langle y_{-j}, b_j\rangle,$$

where $\langle x_{-j}, a_j\rangle = \langle x_1, \ldots, x_{j-1}, a_j, x_{j+1}, \ldots, x_m\rangle$. The following example establishes that in our setting preferences are not separable.

Example 5.6. Let $\mathcal{F}$ be the class of homogeneous linear functions over $\mathbb{R}$, and define $S_1 = \{(\langle 1, 1\rangle, 0)\}$. Then agent 1 prefers $\langle -1, 1\rangle$ to $\langle -1, 2\rangle$, but also prefers $\langle -2, 2\rangle$ to $\langle -2, 1\rangle$.

Border and Jordan [25] investigate a setting where the set of alternatives is $\mathbb{R}^k$. They give possibility results for the case when preferences are star-shaped and separable. On the other hand, when $k \geqslant 2$ and the separability criterion is slightly relaxed, in a way which we will not elaborate on here, then any truthful social choice function must necessarily be dictatorial.

Border and Jordan's results also require surjectivity: the social choice function has to be onto $\mathbb{R}^k$.[5] While this is a severe restriction in general, it is in fact very natural in our context. If all agents have values consistent with some function f, then the mechanism can have a bounded efficiency ratio only if its output is the function f (indeed, f has loss 0, while any other function has strictly positive loss). Therefore, any approximately efficient mechanism must be surjective.

The above discussion leads us to believe that there is no truthful approximation mechanism for homogeneous linear functions over $\mathbb{R}^k$ for any $k \geqslant 2$. The following conjecture formalizes this statement.

Conjecture 5.7. *Let $\mathcal{F}$ be the class of homogeneous linear functions over $\mathbb{R}^k$, $k \geqslant 2$, and assume that $m = |S_i| \geqslant 3$. Then any mechanism that is strategyproof (in ex-post Nash equilibrium) and surjective must be a dictatorship.*

Conceivably, dictatorship would be an acceptable solution if it could guarantee approximate efficiency. A simple example shows that unfortunately this is not the case.

Example 5.8. Consider the class of homogeneous linear functions over $\mathbb{R}^2$, $N = \{1, 2\}$. Let $S_1 = \{(\langle 0, 1\rangle, 0), (\langle 0 + \epsilon, 1\rangle, 0)\}$ and $S_2 = \{(\langle 1, 1\rangle, 1), (\langle 1 + \epsilon, 1\rangle, 1)\}$ for some $\epsilon > 0$. Any dictatorship has an empirical risk of $1/2$. On the other hand, the function $f(x_1, x_2) = x_1$ has empirical risk $\epsilon/2$. The efficiency ratio increases arbitrarily as ϵ decreases.

6. Arbitrary distributions over the sample

In Section 5 we established several positive results in the setting where each agent cares about a uniform distribution on his portion of a global training set. In this section we extend these results to the general regression learning setting defined in Section 3. More formally, the extent to which agent $i \in N$ cares about each point in $\mathcal{X}$ will now be determined by the distribution function ρ_i, and agent i controls the labels of a finite set of points sampled according to ρ_i. Our strategy in this section will consist of two steps. First, we want to show that under standard assumptions on the hypothesis class $\mathcal{F}$ and the number m of samples, each agent's empirical risk on the training set S_i estimates his real risk according to ρ_i. Second, we intend to establish that, as a consequence, our incentive compatibility results are not significantly weakened when we move to the general setting.

[5] Border and Jordan [25] originally required unanimity, but their theorems can be reformulated using surjectivity [26].

Abstractly, let $\mathcal{D}$ be a probability distribution on $\mathcal{X}$ and let $\mathcal{G}$ be a class of real-valued functions from $\mathcal{X}$ to $[0, C]$. We would like to prove that for any $\epsilon > 0$ and $\delta > 0$ there exists $m \in \mathbb{N}$ such that, if $X_1, \ldots, X_m$ are sampled i.i.d. according to $\mathcal{D}$,

$$\Pr\left(\text{for all } g \in \mathcal{G},\ \left|\mathbb{E}_{X\sim\mathcal{D}}\big[g(X)\big] - \frac{1}{m}\sum_{i=1}^{m} g(X_i)\right| \leqslant \epsilon\right) \geqslant 1 - \delta. \tag{19}$$

To establish this bound, we use standard *uniform convergence* arguments. A specific technique is to show that the hypothesis class $\mathcal{G}$ has bounded complexity. The complexity of $\mathcal{G}$ can be measured in various different ways, for example using the pseudo-dimension [27,28], an extension of the well-known VC-dimension to real-valued hypothesis classes, or the Rademacher complexity [29]. If the pseudo-dimension of $\mathcal{G}$ is bounded by a constant, or if the Rademacher complexity of $\mathcal{G}$ with respect to an m-point sample is $O(\sqrt{m})$, then there indeed exists m such that (19) holds.

More formally, assume that the hypothesis class $\mathcal{F}$ has bounded complexity, choose $\epsilon > 0$, $\delta > 0$, and consider a sample S_i of size $m = \Theta(\log(1/\delta)/\epsilon^2)$ drawn i.i.d. from the distribution ρ_i of any agent $i \in N$. Then we have that

$$\Pr\big(\text{for all } f \in \mathcal{F},\ \big|R_i(f) - \hat{R}(f, S_i)\big| \leqslant \epsilon\big) \geqslant 1 - \delta. \tag{20}$$

In particular, we want the events in (20) to hold simultaneously for all $i \in N$, i.e.,

$$\text{for all } f \in \mathcal{F}, \quad \big|R_N(f) - \hat{R}(f, S)\big| \leqslant \epsilon. \tag{21}$$

Using the union bound, this is the case with probability at least $1 - n\delta$.

We now turn to incentive compatibility. The following theorem implies that mechanisms which do well in the setting of Section 5 are also good, but slightly less so, when arbitrary distributions are allowed. Specifically, given a training set satisfying (20) for all agents, a mechanism that is strategyproof in the setting of Section 5 becomes ϵ-strategyproof, i.e., no agent can gain more than ϵ by lying, no matter what the other agents do. Analogously, a group strategyproof mechanism for the setting of Section 5 becomes ϵ-group strategyproof, i.e., there exists an agent in the coalition that gains less than ϵ. Furthermore, efficiency is preserved up to an additive factor of ϵ. We wish to point out that ϵ-equilibrium is a well-established solution concept, the underlying assumption being that agents would not bother to lie if they were to gain an amount as small as ϵ. This concept is particularly appealing when one recalls that ϵ can be chosen to be arbitrarily small.

Theorem 6.1. *Let $\mathcal{F}$ be a hypothesis class, ℓ some loss function, and $S = \biguplus S_i$ a training set such that for all $f \in \mathcal{F}$ and $i \in N$, $|R_i(f) - \hat{R}(f, S_i)| \leqslant \epsilon/2$, and $|R_N(f) - \hat{R}(f, S)| \leqslant \epsilon/2$. Let M be a mechanism with or without payments.*

1. *If M is (group) strategyproof under the assumption that each agent's cost is $\hat{R}(\tilde{f}, S_i)$, then M is ϵ-(group) strategyproof in the general regression setting.*
2. *If M is α-efficient under the assumption that the mechanism's goal is to minimize $\hat{R}(\tilde{f}, S)$, $M(S) = \tilde{f}$, then $R_N(\tilde{f}) \leqslant \alpha \cdot \arg\min_{f\in\mathcal{F}} R_N(f) + \epsilon$.*

Proof. We will only prove the first part of the theorem, and only for (individual) strategyproofness. Group strategyproofness as well as the second part of the theorem follow from similar arguments.

Let $i \in N$, and let $\tilde{u}_i(\tilde{S}_i)$ be the utility of agent i when $\tilde{S}$ is reported and assuming a uniform distribution over S_i. Denoting by $\tilde{f}$ the function returned by M given $\tilde{S}$, we have

$$\tilde{u}_i(\tilde{S}) = -\hat{R}(\tilde{f}, S_i) + p_i(\tilde{S}),$$

where S_i is the training data of agent i with the true labels set by o_i. If M is a mechanism without payments, p_i is the constant zero function. Since M is strategyproof for the uniform distribution, $\tilde{u}_i(S_i, \tilde{S}_{-i}) \geqslant \tilde{u}_i(\hat{S}_i, \tilde{S}_{-i})$ holds for all $\hat{S}_i$.

On the other hand, let u_i denote agent i's utility function with respect to distribution ρ_i, i.e.,

$$u_i(\tilde{S}) = -R_i(\tilde{f}) + p_i(\tilde{S}),$$

where $\tilde{f}$ is as above. Then, $|u_i(\tilde{S}) - \tilde{u}_i(\tilde{S})| = |R_i(\tilde{f}) - \hat{R}(\tilde{f}, S_i)|$. By assumption, this expression is bounded by $\epsilon/2$. Similarly, with respect to i's true values S_i, if $M(S_i, \tilde{S}_{-i}) = \hat{f}$, then

$$\big|u_i(S_i, \tilde{S}_{-i}) - \tilde{u}_i(S_i, \tilde{S}_{-i})\big| = \big|R_i(\hat{f}) - \hat{R}(\hat{f}, S_i)\big| \leqslant \epsilon/2.$$

It follows that for any $\tilde{S}$,

$$u_i(\tilde{S}) - u_i(S_i, \tilde{S}_{-i}) \leqslant \left(\tilde{u}_i(\tilde{S}) + \frac{\epsilon}{2}\right) - \left(\tilde{u}_i(S_i, \tilde{S}_{-i}) - \frac{\epsilon}{2}\right) \leqslant \epsilon. \qquad \square$$

As discussed above, the conditions of Theorem 6.1 are satisfied with probability $1-\delta$ when $\mathcal{F}$ has bounded dimension and $m = \Theta(\log(1/\delta)/\epsilon^2)$. As the latter expression depends logarithmically on $1/\delta$, the sample size only needs to be increased by an additive factor of $\Theta(\log(n)/\epsilon^2)$ to achieve the stronger requirement of (21).

Let us examine how Theorem 6.1 applies to our positive results. Since ERM with VCG payments is strategyproof and efficient under uniform distributions over the samples, we obtain ϵ-strategyproofness and efficiency up to an additive factor of ϵ when it is used in the general learning game, i.e., with arbitrary distributions. This holds for any loss function ℓ. The project-and-fit mechanism is ϵ-group strategyproof in the learning game when $\mathcal{F}$ is the class of constant functions or of homogeneous linear functions over $\mathbb{R}$, and 3-efficient up to an additive factor of ϵ. This is true only for the absolute loss function.

7. Discussion

In this paper, we have studied mechanisms for a general regression learning framework involving multiple strategic agents. In the case where each agent controls one point, we have obtained a strong and surprising characterization of the truthfulness of ERM. When the absolute loss function is used, ERM is group strategyproof. On the other hand, ERM is not strategyproof for any loss function that is superlinear in a certain well-defined way. This particularly holds for the popular squared loss function. In the general learning setting, we have established the following result: For any $\epsilon, \delta > 0$, given a large enough training set, and with probability $1-\delta$, ERM with VCG payments is efficient up to an additive factor of ϵ, and ϵ-strategyproof. We have also obtained limited positive results for the case when payments are disallowed, namely an algorithm that is ϵ-group strategyproof and 3-efficient up to an additive factor of ϵ for constant functions over $\mathbb{R}^k$, $k \in \mathbb{N}$, and for homogeneous linear functions over $\mathbb{R}$. We gave a matching lower bound, which also applies to multi-dimensional homogeneous linear functions. The number of samples required by the aforementioned algorithms depends on the combinatorial richness of the hypothesis space $\mathcal{F}$, but differs only by an additive factor of $\Theta(\log(n)/\epsilon^2)$ from that in the traditional regression learning setting without strategic agents. Since $\mathcal{F}$ can be assumed to be learnable in general, this factor is not very significant.

At the moment there is virtually no other work on incentives in machine learning, many exciting directions for future work exist. While regression learning constitutes an important area of machine learning with numerous applications, adapting our framework for studying incentives in classification or in unsupervised settings will certainly prove interesting as well. In classification, each point of the input space is assigned one of two labels, either $+1$ or -1. ERM is trivially incentive compatible in classification when each agent controls only a single point. The situation again becomes complicated when agents control multiple points. In addition, we have not considered settings where ERM is computationally intractable. Just like in general algorithmic mechanism design, VCG is bound to fail in this case. It is an open question whether one can simultaneously achieve tractability, approximate efficiency, and (approximate) incentive compatibility.

Several interesting questions follow directly from our work. The one we are most interested in is settling Conjecture 5.7: are there incentive compatible and approximately efficient mechanisms without payments for homogeneous linear functions? Do such mechanisms exist for other interesting hypothesis classes? These questions are closely related to general questions about the existence of incentive compatible and non-dictatorial mechanisms, and have implications way beyond the scope of machine learning and computer science.

Acknowledgments

We thank David Parkes, Bezalel Peleg, and Jeff Rosenschein for helpful discussions, and Yishay Mansour for pointing us to the work of Perote and Perote-Peña [8]. We are also grateful to an anonymous referee for valuable comments.

References

[1] F. Caro, J. Gallien, Inventory management of a fast-fashion retail network, Oper. Res., doi:10.1287/opre.1090.0698, forthcoming.
[2] F. Caro, J. Gallien, M.D. Miranda, J.C. Torralbo, J.M.C. Corras, M.M. Vazquez, J.A.R. Calamonte, J. Correa, Zara uses operations research to reengineer its global distribution process, Interfaces 40 (1) (2010) 71–84.
[3] N. Littlestone, Redundant noisy attributes, attribute errors, and linear-threshold learning using Winnow, in: Proceedings of the 4th Annual Workshop on Computational Learning Theory (COLT), 1991, pp. 147–156.
[4] S.A. Goldman, R.H. Sloan, Can PAC learning algorithms tolerate random attribute noise?, Algorithmica 14 (1) (1995) 70–84.
[5] M. Kearns, M. Li, Learning in the presence of malicious errors, SIAM J. Comput. 22 (4) (1993) 807–837.
[6] N.H. Bshouty, N. Eiron, E. Kushilevitz, PAC learning with nasty noise, Theoret. Comput. Sci. 288 (2) (2002) 255–275.
[7] M. Barreno, B. Nelson, R. Sears, A.D. Joseph, J.D. Tygar, Can machine learning be secure?, in: Proceedings of the 1st ACM Symposium on Information, Computer and Communications Security (ASIACCS), ACM Press, 2006, pp. 16–25.
[8] J. Perote, J. Perote-Peña, Strategy-proof estimators for simple regression, Math. Social Sci. 47 (2004) 153–176.
[9] N. Nisan, A. Ronen, Algorithmic mechanism design, Games Econom. Behav. 35 (1–2) (2001) 166–196.
[10] D. Lehmann, L.I. O'Callaghan, Y. Shoham, Truth revelation in rapid, approximately efficient combinatorial auctions, J. ACM 49 (5) (2002) 577–602.
[11] S. Dobzinski, N. Nisan, M. Schapira, Truthful randomized mechanisms for combinatorial auctions, in: Proceedings of the 38th Annual ACM Symposium on Theory of Computing (STOC), 2006, pp. 644–652.
[12] M.-F. Balcan, A. Blum, J.D. Hartline, Y. Mansour, Reducing mechanism design to algorithm design via machine learning, J. Comput. System Sci. 74 (8) (2008) 1245–1270.
[13] G. Kalai, Learnability and rationality of choice, J. Econom. Theory 113 (1) (2003) 104–117.

[14] A.D. Procaccia, A. Zohar, Y. Peleg, J.S. Rosenschein, The learnability of voting rules, Artificial Intelligence 173 (12–13) (2009) 1133–1149.
[15] N. Nisan, Introduction to mechanism design (for computer scientists), in: N. Nisan, T. Roughgarden, E. Tardos, V. Vazirani (Eds.), Algorithmic Game Theory, chap. 9, Cambridge University Press, 2007, pp. 209–265.
[16] W. Vickrey, Counter speculation, auctions, and competitive sealed tenders, J. Finance 16 (1) (1961) 8–37.
[17] E.H. Clarke, Multipart pricing of public goods, Public Choice 11 (1971) 17–33.
[18] T. Groves, Incentives in teams, Econometrica 41 (1973) 617–631.
[19] M. Rothkopf, Thirteen reasons the Vickrey–Clarke–Groves process is not practical, Oper. Res. 55 (2) (2007) 191–197.
[20] A. Mas-Colell, M.D. Whinston, J.R. Green, Microeconomic Theory, Oxford University Press, 1995.
[21] H. Moulin, Generalized Condorcet-winners for single peaked and single-plateau preferences, Soc. Choice Welf. 1 (2) (1984) 127–147.
[22] J. Shawe-Taylor, N. Cristianini, Support Vector Machines and Other Kernel Based Learning Methods, Cambridge University Press, 2000.
[23] A. Gibbard, Manipulation of voting schemes, Econometrica 41 (1973) 587–602.
[24] M. Satterthwaite, Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions, J. Econom. Theory 10 (1975) 187–217.
[25] K. Border, J. Jordan, Straightforward elections, unanimity and phantom voters, Rev. Econom. Stud. 50 (1983) 153–170.
[26] Y. Sprumont, Strategyproof collective choice in economic and political environments, Can. J. Econ. 28 (1) (1995) 68–107.
[27] D. Pollard, Convergence of Stochastic Processes, Springer-Verlag, 1984.
[28] D. Haussler, Decision theoretic generalization of the PAC model for neural net and other learning applications, Inform. and Comput. 100 (1) (1992) 78–150.
[29] P.L. Bartlett, S. Mendelson, Rademacher and Gaussian complexities: Risk bounds and structural results, J. Mach. Learn. Res. 3 (2003) 463–482.